Dissertation presented to Instituto Tecnológico de Aeronáutica and Universidade Federal de São Paulo, in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program of Industrial Engineering, Field of Operations Research.

Guilherme Ribeiro da Silva

FEATURE SELECTION FOR CHARACTERIZATION OF CONTINUOUS OPTIMIZATION FUNCTIONS

Dissertation approved in its final version by signatories below:

Prof. Dr. Rodrigo A. Scarpel
Advisor

Prof. Dr. Ana Carolina Lorena
Co-advisor

São José dos Campos, SP - Brazil
2019

Cataloging-in Publication Data
Documentation and Information Division

Ribeiro da Silva, Guilherme
Feature Selection for Characterization of Continuous Optimization Functions / Guilherme Ribeiro da Silva.
São José dos Campos, 2019.
129f.

Dissertation of Master of Science – Course of Industrial Engineering. Area of Operations Research – Instituto Tecnológico de Aeronáutica and Universidade Federal de São Paulo, 2019. Advisor: Prof. Dr. Rodrigo A. Scarpel. Co-advisor: Prof. Dr. Ana Carolina Lorena.

1. Feature Selection. 2. Continuous Optimization. 3. Dimensionality Reduction. 4. Black Box Optimization Benchmark. I. Instituto Tecnológico de Aeronáutica. II. Universidade Federal de São Paulo III. Title.

BIBLIOGRAPHIC REFERENCE

RIBEIRO DA SILVA, Guilherme . **Feature Selection for Characterization of Continuous Optimization Functions**. 2019. 129f. Dissertation of Master of Science – Instituto Tecnológico de Aeronáutica and Universidade Federal de São Paulo, São José dos Campos.

CESSION OF RIGHTS

Guilherme Ribeiro da Silva
Av. São João, 2655
12.242-000 – São José dos Campos–SP

FEATURE SELECTION FOR CHARACTERIZATION OF CONTINUOUS OPTIMIZATION FUNCTIONS

Guilherme Ribeiro da Silva

Thesis Committee Composition and Signature:

Prof. Dr.	Rafael Duarte Coelho dos Santos	External Member	-	INPE
Prof. Dr.	Rodrigo A. Scarpel	Advisor	-	ITA
Prof. Dr.	Ana Carolina Lorena	Co-advisor	-	ITA
Prof. Dr.	Antonio A. Chaves	Internal Member	-	UNIFESP
Prof. Dr.	Marcos Gonçalves Quiles	Internal Memmber	-	UNIFESP

ITA/UNIFESP

To all my friends.

Acknowledgments

First, I would like to thank my supervisor Dr. Rodrigo A. Scarpel and Co-Supervisor Dr. Ana Carolina Lorena for all the support, knowledge and patience given to me during my time as a student. I truly appreciate all the guidance as it paved me the path to become a better researcher.

Secondly, I would like to thank my friends Rosana, Carla and Rafaela for the moments we shared as graduate students when times turned to prove difficult.

Lastly, a special thanks to my family, who gave me all the support necessary so I could focus entirely on my studies and personal development.

"All that is gold does not glitter,
Not all those who wander are lost;
The old that is strong does not wither,
Deep roots are not reached by the frost."
— J.R.R Tolkien

Resumo

O tema de Seleção Automática de Algoritmos tem recebido crescente atenção nos últimos anos. Técnicas de Aprendizagem de Máquinas são capazes de predizer com alta precisão o melhor conjunto de algoritmos para uma dada instância de uma função. No presente trabalho, são estabelecidos as etapas necessárias para caracterização de funções contínuas. Com uma abordagem de Redução de Dimensionalidade, apenas os atributos que melhor preservem a informação do conjunto de dados serão selecionados. Então, o conjunto de atributos foi comparado com métricas já estabelecidas na literatura e os resultados analisados.

Abstract

The field of automatic algorithm selection has received increased attention in the past years. Machine Learning techniques are now able to predict with high accuracy the best set of algorithms for a given problem instance. In this present dissertation, we establish the steps for the selection of features for characterization of continuous functions. Via a Dimensionality Reduction approach, only the features that best preserve the information of the dataset are selected. Then, the selected features were compared to already established sets of metrics from the literature and the results are analyzed.

List of Figures

List of Tables

List of Abbreviations and Acronyms

ELA	Exploratory Landscape Analysis
DR	Dimensionality Reduction
BBOB	Black Box Optimization Benchmark
ASP	Algorithm Selction Problem
GA	Genetic Algorithm
ML	Machine Learning
LS	Length Scale
PCA	Principal Component Analysis

Contents

1 Introduction

1.1 Motivation

Researchers have often come across the goal of optimizing complex and hard objective functions. Many examples of such functions are seen in the fields of Engineering, Bioinformatics, Finances, and Computer Science, only to name a few. Therefore, there is a need for algorithms that are capable of providing good solutions for increasingly complex problems (GOGNA; TAYAL, 2013).

Meanwhile, the field of optimization has seen the number of algorithms increase exponentially throughout the years. Operations Research deals specifically with solving problems within distinct areas such as Networks, Logistics, Manufacturing, for example, (GONZÁLEZ *et al.*, 2014) of different nature such as linear, nonlinear, combinatorial or integer. With a specific problem of a given category, researchers may be surrounded by a plethora of algorithms to select from.

In this sense, not only researchers come across the difficult task of finding a possible optimal solution to their problem, but may find it difficult for selecting a suitable algorithm. Besides, each algorithm possesses a set of tuning parameters that have a great influence on the final achieved performance. In the hope of addressing a specific problem with the correct technique, some works have been developed to characterize optimization functions. Exploratory Landscape Analysis (ELA) is a set of metrics employed to obtain the properties of an unknown optimization problem (MERSMANN *et al.*, 2011), therefore, with such acquired information better algorithms can be designed and later employed into more problem-specific situations (GOLLE, 2011).

Many applications of ELA are found in the literature. For example, (GOLLE, 2011) used autocorrelation and fitness-distance correlation metrics for understanding a combinatorial car sequencing problem. (MERZ, 2004) used a set of fitness landscape techniques to understand the performance of a class of metaheuristic algorithms called Memetic Algorithms. (NOZ *et al.*, 2015) proposed an improved version of an information content metric and (MORGAN; GALLAGHER, 2017) proposed a metric capable of efficiently differentiating between a group of highly modal optimization functions considered hard.

Therefore, an area of research that recently gained attraction is the field of Algorithm Selection Problem (ASP) (RICE, 1976). Dealing with the specific problem of selecting the most suitable technique from a portfolio (or bag of algorithms), much of the research deals with linking the underlying function characteristics to algorithm performance or the most suitable algorithm. Besides, much of the increased popularity of the ASP is due to the adoption of powerful techniques from Machine Learning and Pattern Recognition fields. In (MUÑOZ *et al.*, 2012), (BISCHL *et al.*, 2012), (KERSCHKE; TRAUTMANN, 2019) and (GONÇALVES, 2018), the authors performed work in algorithm selection for black-box optimization functions. They mapped different ELA metrics to algorithm performance in a known benchmark of functions for continuous optimization. Though they have enjoyed success in creating a mapping model, each of the works used a different set of ELA metrics, which measure a different aspect of the functions. This, therefore, may have had an influence in properly characterizing different functions and in the final performance of the mapping model between ELA metrics and algorithm performance.

1.2 Objective

The main objective of the present dissertation is to explore the selection of ELA metrics for the characterization of continuous optimization problems. Because each metric measures a specific characteristic from categories such as smoothness, modality, and neutrality, for example, by carefully selecting different features, one may be able to characterize and classify different optimization functions according to their intrinsic characteristics. This, in turn, is useful for the optimization community as it provides a means for properly understanding what differentiates one function type from another. Moreover, this line of research can lead to the development of new benchmark functions as well as give clarity to what makes an optimization problem to be considered hard. Researchers in the field of ASP may also benefit from the present work. By using fewer but meaningful ELA metrics, a mapping model increases its prediction power, such as evidenced in works related to the importance of feature selection in Machine Learning as in (CHANDRASHEKAR; SAHIN, 2014).

Therefore, the present work objective can be resumed as:

The selection of features based in Exploratory Landscape Analysis for characterizing continuous functions. Where the main areas which benefit from this approach are:

1. A set of carefully chosen ELA metrics are useful for differentiating and comparing continuous optimization functions.

2. A set of carefully chosen ELA metrics are useful for the initial steps of building an algorithm recommender system.

In the present work, the selection of features shall be performed via a Dimensionality Reduction (DR) approach, or more specifically, Manifold Learning (HUO; SMITH, 2008). The overall idea is to represent a high dimensional data set (composed of many different features) in a lower-dimensional space. Only the features which best preserve the original structure of the dataset shall be considered. In this manner, a metric that measures the information loss between the original higher-dimensional dataset and its projection in a lower dimension shall be used. In (GRACIA *et al.*, 2014), the authors compare different quality loss criteria and dimensionality reduction techniques to choose the most suitable criteria and technique according to a dataset underlying structure. The interested reader is referred to (GRACIA *et al.*, 2014) for further details.

Specifically, the Co-Ranking quality criteria (LEE; VERLEYSEN, 2007) is adopted to assess the preservation of the dataset information structure in a lower-dimensional space. Although the use of DR techniques in conjunction with ELA metrics is not new, the way it has been used in similar works differs fundamentally from what is proposed in this work. For instance, in (MORGAN; GALLAGHER, 2017), the authors use a nonlinear DR technique called t-SNE to compare their proposed ELA metric (Length Scale) compared to other standard ELA metrics in characterizing functions of the BBOB09 benchmark set (HANSEN *et al.*, 2009). The authors were indeed able to produce 2D plots which visually depicted how the two different sets of ELA metrics portrayed the dissimilarity of functions in a reduced space, but their approach fails to specifically show how well the 2D embedded space reflects the structure of true, higher dimensional dataset by means of an appropriate information loss metric. Similarly, in (MUÑOZ; SMITH-MILES, 2015), the authors employ PCA to embed a set of ELA metrics in a lower-dimensional 2D space. They analyze the distribution of each function type in a 2D plot and assess the explained variance from the original dataset in each principal component.

Differently, in this work, DR techniques shall be used beyond depicting continuous functions in a 2D space and assessing their dissimilarity. A specific quality criterion (information loss) will be used to guide the selection of features; that is, only the features which best preserve the information of the original dataset shall be selected. So far, the author has not found any similar work in the literature.

1.3 Problem Statement and Challenges

The selection of features for characterizing continuous functions involves addressing some challenges. First, an overall set of features must be assessed for feature selection

to take place. Because the field of ELA has been studied since the early 1990's (JONES; FORREST, 1995), a vast myriad of metrics are available to be selected. Secondly, one should carefully choose a set of representative continuous functions that possess different characteristics and which could later generalize the behavior of unseen continuous functions, such as in practical real-world problems. Questions related to DR algorithms are also important. Some DR techniques are not able to cope with nonlinear structures in the dataset. Therefore, the choice of suitable algorithms for DR is paramount for a truthful representation of a data set in a lower-dimensional space. Hence, the following issues are posed:

- Which DR algorithms shall be used to perform dimensionality reduction?
- In which continuous functions should the DR algorithms be tested?
- What is the set of ELA metrics that selection must take place?
- How to ensure the lower-dimensional representation captures the overall structure of the higher dimensional data?

A quick search in the literature of DR algorithms shows an explosion in the number of different methods. DR techniques have been widely used in the field of Machine Learning for assessing the intrinsic characteristics of the data. For instance, t-SNE (MAATEN; HINTON, 2008), kernel PCA (HOFFMANN, 2007), Isomap (SILVA; TENENBAUM, 2003) have been popular choices for nonlinear dimensionality reduction, whereas PCA and MDS have succeeded in linear dimensionality reduction. For a thorough account of different DR techniques the interested reader is referred to (CUNNINGHAM; GHAHRAMANI, 2015) and (LEE; VERLEYSEN, 2007).

Also, within the scope of optimization of continuous functions, many researchers produced benchmarks functions for algorithm comparison. Known benchmarks are CEC 2005 (SUGANTHAN *et al.*, 2005), BBOB09 (HANSEN *et al.*, 2009) and ZDT (ZITZLER *et al.*, 2000). The idea of a benchmark is to provide a list of functions that possesses different characteristics that could be explored for experimental purposes. These fundamental characteristics such as ruggedness, valleys or curvatures in high-dimensional space are believed to make the search for an optimal solution more or less complex. In the present work, we focus on the benchmark functions of BBOB09 (HANSEN *et al.*, 2009), where the authors selected 24 functions divided into characteristics that shall be detailed later in Chapter 2: Literature Review.

As mentioned the research community has produced a wide number of ELA metrics, each of which characterizes a specific aspect of a function. To address the problem of collecting many metrics under a single database, (KERSCHKE, 2017) developed a package

joining a wide range of ELA metrics. This package has been used for characterizing different functions and as well as for purposes of automated algorithm selection as in (KERSCHKE; TRAUTMANN, 2019).

Since one of the objectives of optimizing a function is to give the best set of variables which minimizes (or maximizes) a given problem, one natural way to characterize algorithms is to compare how close they get to a known previous or standard solution. Here, one may then see the usefulness of crafted developed benchmark functions.

Finally, because the overall goal of the present work is to perform feature selection based in DR techniques, a suitable metric for measuring the information loss from a higher to lower space shall be used. (LEE; VERLEYSEN, 2007) proposed a Co-Ranking framework that provides a different number of relevant quality loss criteria that shall be further explained in Chapter 2: Literature Review.

1.4 Research Directions

Initially, a set of ELA metrics found in (KERSCHKE, 2017) are used to characterize each selected function in the BBOB09 benchmark (HANSEN *et al.*, 2009). Afterward, DR is used to project (embed) the calculated metrics into a lower-dimensional space to provide information about the structure of the data set. This analysis shows how the similarity or distance among points representing functions forms a cloud of visible patterns. Then, after a feature (metric) selection procedure, only a small subset of the original metrics are selected.

In the second step, the selected set of features are compared to related works. That is, the 2D lower dimensional space produced by the proposed sets are compared visually to the embedded 2D plot of similar sets of ELA metrics. Then a discussion will be made on the nature of the metrics and how the similarity between function changes according to each set of metrics. The Co-Ranking Framework shall be used to quantitavely assess the information preservation of the manifold when performing dimensionality reduction.

1.5 Dissertation Contribution

As mentioned, the optimization community, as well as researchers from the field of Algorithm Selection Problem may benefit from the results of this dissertation. The main contributions of the present work are resumed as follows:

1. **Proposal of a new set of ELA metrics**: The new set of ELA metrics are useful for developing models linking function characteristics and algorithm performance as

in the ASP, or for exploration of functions characteristics as in the case of research in ELA metrics.

2. **Usage of a DR criteria framework for assessing the quality of embedded (projected) function characteristics**: Though the usage of ELA metrics in conjunction with DR techniques is not new, in this thesis an alternative way of assessing the quality of embedded ELA metrics data is proposed. In (MORGAN; GALLAGHER, 2017), the authors used the cost function of the t-SNE DR for assessing the quality of the embedding. The use of the Co-Ranking framework offers a specific set of metrics developed specifically for analyzing embedded data. For instance, it is possible to answer how different is the quality of local or near data points (functions) in comparison to the global structure of the embedded data. This framework thus paves the way for the discussion of which DR technique should be employed, in how well the 2D plot reflects a higher dimensional dataset information structure locally or globally.

3. **A procedure for selecting ELA metrics**: Through the combination of an optimization algorithm and the Co-Ranking framework a set of features (ELA metrics) to characterize functions are selected. Related works in the literature of ELA metrics have so far qualitatively selected metrics, that is, selecting metrics related to some characteristics without considering further quantitative aspects.

4. **A Comparison of the different sets of ELA metrics**: A comparison of the proposed set of ELA metrics in the literature may reveal interesting information about the BBOB09 set of functions, the capabilities of the DR techniques and the strengths and weaknesses of metrics.

1.6 Text Organization

The present work is organized as follows: In Chapter 2: Literature Review, the main theoretical aspects of the dissertation are reviewed. First, a set of benchmark functions for continuous optimization are introduced and then the main benchmark, that is, the BBOB09 (HANSEN *et al.*, 2009). Then the theory of Dimensionality Reduction is introduced as well as the formulation of the t-SNE DR technique. A brief introduction of the Algorithm Selection Problem is provided to contextualize the reader. Finally, the chapter ends with a summary of related works.

In Chapter 3: Materials and Methods the procedural experiments are detailed. After a quick introduction of the process of feature selection of ELA metrics, the justification for conducting this line of research is retaken. This justification, in turn, shall serve as a

guideline to detail each step in the experimental setup in which the present thesis is based on.

In Chapter 4: Result Analysis, the final results are discussed. The selected set of ELA metrics proposed by the present work are compared with the already established set of ELA metrics from related works. First, a visual exploratory analysis is performed to depict the range of assumed values for each ELA metric for all function categories. In the second step, the comparison shall be made using the 2D embedded scatter plot, the standard way of depicting the BBOB09 benchmark of continuous functions through ELA metrics. Finally, each set of ELA metric shall inspected using the Co-Ranking Framework. In this way, the information preservation produced by each set of ELA metrics will be quantified.

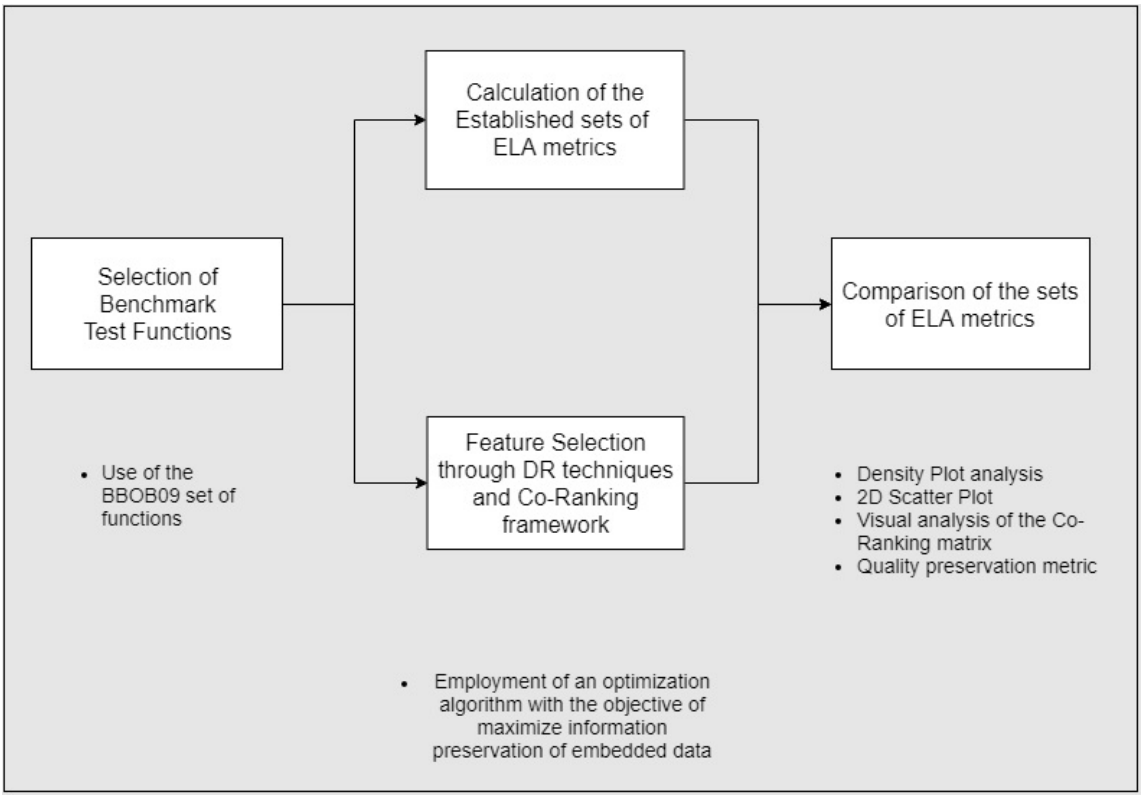


FIGURE 1.1 – Workflow of present work.

2 Literature Review

2.1 Preliminaries

In this section, the necessary theory behind the present work is provided. First, the Black Box Optimization Benchmark (BBOB09) (HANSEN *et al.*, 2009) is detailed as well as the need for a benchmark of continuous functions. Then, different aspects of the theory of Exploratory Landscape Analysis (ELA) are introduced. The theory is relevant because it provides actual relevant guidance on characteristics a researcher shall observe when studying a function's underlying structure. Also, because one single facet or metric is incapable of fully describing a function structure at its entirety, different criteria such as smoothness, ruggedness, neutrality and so on are explored. As stated in Chapter 1: Introduction, the objective of the present dissertation is the selection of features (ELA metrics) that preserve the most the dataset structure.

In the second step, the theory of Dimensionality Reduction (DR) is presented. The theory on DR will serve as a step to check whether different proposed ELA metrics are capable of describing and classifying function categories. By depicting functions as points in a lower-dimensional space such as 2D, one may notice how similar or dissimilar functions are by just checking the distance between these points in a plot. Other interesting stylized facts that may be studied are the emergence of clouds of points or patterns, which could be useful for separating functions into different subgroups or subclasses. One natural question in the use of DR techniques is how truthfully does the embedded dataset reflects the real structure or patterns of a higher, unseen dimensional space. To answer these questions researchers developed different techniques to measure the information loss of an embedded dataset. In this work, a metric based on the Co-Ranking matrix (LEE; VERLEYSEN, 2009) is adopted and further explained later in this chapter.

Finally, the chapter ends with a discussion on related works and similarities and divergences to the present work are pointed out.

2.2 A Benchmark of Continuous Functions

2.2.1 Introduction to Benchmarks

The Black-Box Optimization Benchmark (BBOB09) (HANSEN *et al.*, 2009) is comprised of 24 different types of functions. The authors intended to create a set that presents different topological signatures that would make it easier to study an optimization algorithm behavior - success or failure in finding an optimal where the underlying characteristics of a function are understood a priori. The 24 functions are defined in 5 distinct categories: Separable functions, Functions with low or moderate conditioning, Functions with high conditioning and unimodal, Multi-modal functions with adequate global structure and Multi-modal functions with weak global structure. The BBOB set was used extensively yearly in a competition scheme at the Genetic and Evolutionary Computation Conference (GECCO), to compare novel and state of the art algorithms developed by researchers. In (HANSEN *et al.*, 2010), the authors compare the results of 31 optimization algorithms with the BBOB09 set functions. Also, the results of the yearly competition are available at (COCO, 2019).

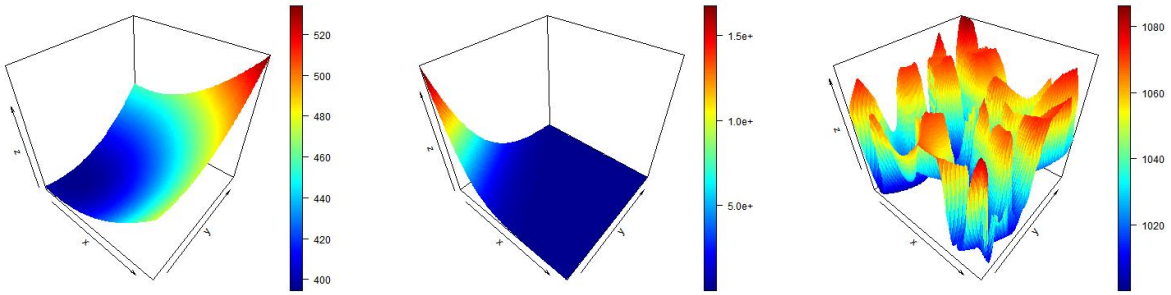
Similarly, the Congress of Evolutionary Computation (SUGANTHAN *et al.*, 2005) has also produced a set of functions for algorithm comparison within a competition scheme. However, differently than the aforementioned benchmark is the Zitzler set of functions (ZITZLER *et al.*, 2000). Whereas the CEC and BBOB comprise of a single objective optimization test functions, the Zitzler - or ZDT - serves as a benchmark for multi-objective optimization, that is, the idea isn't in checking how far a produced solution is from a known optimum point but an entire set of points known as the Pareto frontier. The development of new benchmark sets is an active field of research and already established functions are being transformed to produce functions with new topological signatures. For instance, in (QU *et al.*, 2016) the authors extend already established multi-modals functions by applying coordinate shifts and rotation operations.

2.2.2 Characteristics of the BBOB09

It is interesting to note that because these functions are artificially created by rotating and shifting a known generative model in a specified space, a researcher can produce new instances of the same function type for further investigation. For example, for the same configuration such as the number of dimensions and function type, one can study if these different configurations produce similar structures in the space. Table A.1 in Appendix A resumes the theoretical characteristics from the included in the set as given in (HANSEN *et al.*, 2009). The author defines in 6 categories the function properties:

- **Deceptive Functions**: A structure within the landscape that deceives an optimizer to explore a given area.
- **Ill-Conditioning**: For quadratic functions, $f(x) = \frac{1}{2}x^T Hx$, with H symmetric definite positive. For general functions, the authors describe it loosely as the square of the ratio between the largest and small direction of a contour line.
- **Regularity**: The regularity of a given structure across the search space.
- **Separability**: The landscape may be characterized by different subsets of topological properties.
- **Symmetry**: The function possesses symmetry of its landscape with respect to an axis.
- **Target function value to reach**: It reflects the scalar multiplier and its interference in the search of the function global optimal.

In Appendix A the definitions to generate BBOB24 different types are given. Figure 2.1 shows the topologies of three types of BBOB functions, namely BBOB1, BBOB12, and BBOB12. In Appendix B the functions in the set are produced for visual inspection. From figure 2.1 it is noticeable how different functions from the benchmark possess different underlying structures. Therefore the test set opens the research question in how some optimization algorithms might produce quality or poor solutions when the underlying function characteristics are known.



(a) BBOB1 (b) BBOB12 (c) BBOB22

FIGURE 2.1 – Landscape of three BBOB functions. From the author.

2.3 Exploratory Landscape Analysis

2.3.1 Introduction of Function Characteristics and Algorithm Selection

(SMITH-MILES, 2009) draws attention to the need for a better understanding of the characteristics of a class of problems to carry out the selection of an appropriate algorithm. The key to characterizing a function is the use of easy-to-calculate metrics that can capture the different characteristics of the problem at hand.

Thus, the question to be raised is, for certain classes or instances, what are the metrics or characteristics which have relation to the performance of a given algorithm? And, is it possible to design a model that relates the characteristics of a class of problems to the performance of algorithms? The literature presents different metrics that can be used in the evaluation of optimization problems.

(ALPCAN *et al.*, 2014) establishes the use of Shannon's Information and Entropy Theory to define an upper limit to be used as a proxy for the difficulty of an optimization problem. On the other hand, many researchers seek to understand the intrinsic structures of a given class of problems and relate them to the difficulty of finding an optimal solution. This technique is called Fitness Landscape Analysis and covered extensively in (GOLLE, 2011), (MALAN; ENGELBRECHT, 2013), (MERZ, 2004), (WATSON, 2010). Besides, some researchers have evolved the concept of Fitness Landscape for exploratory use, so the structure of the problem is determined quickly for almost instantaneous use, such as in the case of automatic algorithm selection (MERSMANN *et al.*, 2010).

Other problem characterization metrics are based on Complexity Theory (BORENSTEIN; POLI, 2006), mapping of cells based on Markov Chains (KERSCHKE *et al.*, 2014), Length Scale Article (MORGAN; GALLAGHER, 2017), Measures of Dispersion (LUNACEK; WHITLEY, 2006) and others. Therefore, it is valid to understand the different characteristics of optimization problems according to different authors to extract useful metrics for the selection of algorithms that reflect a certain characteristic of an optimization function.

2.3.2 Function Characteristics and Optimization Hardness

Due to the complexity of optimization problems, researchers have concluded that it is difficult to classify a "difficult" optimization problem with the use of only one metric (MALAN; ENGELBRECHT, 2013). In this way, many classifications were created to capture the different characteristics that have a relation with the difficulty of problem optimization. In the following subsections, the main characteristics of functions are described

according to different researchers. It is interesting to note what are the divergences and similarities of the proposed categories of function characteristics.

2.3.2.1 Function Characteristics according to (MALAN; ENGELBRECHT, 2013)

According to (MALAN; ENGELBRECHT, 2013), different characteristics shall be used to describe the behavior of complex functions:

1. **Degree of variable inter-dependency - Epistasis**: It characterizes the dependency of assumed values of a variable into other variables - also known as nonlinear separability. The authors state that when optimization problems possess dependent variables, it is impossible to find the optimal value one variable without compromising another, thus making the search difficult.

2. **Noise**: It is defined as a characteristic that includes uncertainties in the measurements and values of a given function.

3. **Fitness Distribution**: It characterizes the assumed fitness values of the solutions in the entire search space. From this characteristic statistical metrics can be calculated as mean, median, kurtosis, and asymmetry, among others.

4. **Fitness Distribution in the search space**: Similar to fitness distributions of the entire search space, but reduced to a fraction of space. This characteristic measures how heterogeneous a function space can be in different regions of the landscape.

5. **Modality**: It characterizes the number of peaks that a function presents. Researches show that multimodal functions tend to present greater difficulty in being optimized, that is, finding the optimum global solution.

6. **Information and Disinformation**: The authors point out that some problems have an intrinsic structure such that they guide a certain algorithm more easily to a global optimum. In this way, the quantity, but also the quality of information is important, since the information contained in the problem can direct the algorithm in the wrong direction. Also, it is worth noting that this feature depends on the nature of the algorithm, and not just on the structure of the problem.

7. **Global Structure**: The global structure of a problem is related to how points of high fitness values are grouped in the search space. The author points out that even multi-modal functions can be relatively easy to optimize if there is a global structure that groups together points in a manner that can be exploited by algorithms, such as the Rastrigin continuous function type.

8. **Ruggedness and smoothness**: A function is characterized as rough if nearby solutions have many different fitness values. The opposite of a rough function, that is, a smooth function, is a function that has no elevations or heaps.

9. **Neutrality**: Neutrality is present in the solutions space if neighboring solutions have the same fitness value. This characteristic present in different functions increases the difficulty of its solution because an optimization algorithm may behave as if the solution has converged to an optimal solution.

10. **Symmetry**: Symmetric search spaces present solutions with similar fitness values. Problems can be symmetrical with respect to an axis, but also to certain optimal solutions. Researchers have shown that symmetry may harm the performance of certain evolutionary algorithms.

11. **Evolvability/Searchabiliy**: Defined as the ability of an algorithm to produce better fitness solutions than its parents' solutions, this feature was later adapted to encompass the term "search" so as not to restrict the term to just evolutionary algorithms. The ability to search for a problem relates to the ability of an algorithm to move from one region with low fitness to another region whose solutions are better. This characteristic is also dependent on the type of algorithm used.

2.3.2.2 Function Characteristics according to (MERSMANN *et al.*, 2011)

(MERSMANN *et al.*, 2011) also seeks to define characteristics of optimization problems. The authors use metrics based on the here defined characteristics in conjunction with Machine Learnings models. Their objective in the paper was to test whether ELA metrics were capable of classifying the different 24 BBOB function categories.

1. **Multi-Modality**: A characteristic where function presents a different number of local optima within a region of the function space.

2. **Global Structure**: Defined as the structure of points adjacent to local optima in the entire search space.

3. **Separability**: This characteristic allows a function to be partitioned into smaller different problems. In this way, with the reduction of its size, it is believed that a search algorithm may find it easier to find an optimal solution. However, the authors point out this characteristic not always present in functions.

4. **Scalability**: The scalability of a problem is related to the variation of its fitness values according to small adjustments in the values of variables. A problem that presents high scalability has a variable that when suffers small adjustments greatly

impacts the fitness values of the function when compared to the same small adjustment in other variables of the problem.

5. **Homogeneity in the Search Space**: A problem with such characteristics presents similar regions throughout its search space. Homogeneous problems do not have abrupt transitions or phases in their landscape.

6. **Homogeneity of the Basins of Attraction**: It characterizes the width and height of the different basins of attraction of a problem. It is believed that heterogeneous attraction pools increase the difficulty of an optimization problem.

7. **Global-Local Optimal Contrast**: It characterizes the difference between the average fitness values of global solutions to the average fitness values of overall, or local optima solutions. This characteristic attempts to describe whether optimal points of the function are easily recognized.

8. **Plateau**: Characterizes regions of the search space where solutions have equal or very similar fitness values.

2.3.2.3 Function Characteristics according to (ABELL *et al.*, 2013)

(ABELL *et al.*, 2013) simplify the process of characterization of continuous functions and propose only three categories. In each category, statistical metrics are calculated to capture the overall structure of the problem. The authors cite that simplicity in the computation of metrics must be sought so that the excessive computational cost of obtaining metrics removes all the benefits of quickly characterizing a function.

1. **Problem Definition**: Metrics in this category are related to the number of variables and the required precision of the optimization algorithm. The set of these metrics helps to describe the size of the optimization problem.

2. **Optimal Solutions Characteristics**: Metrics in this category describe the distance between optimal local solutions, their fitness values, standard deviation, and density of such solutions in a search space region. Then it with such metrics it is possible to better understand the difficulty algorithms encounter when searching for the optimal global solution.

3. **Random Solutions**: In this category, metrics are calculated to compare characteristics of random solutions in the search space to optimal solutions. In this manner, metrics such as mean and standard deviation of solutions concerning an optimal solution are calculated.

The three sections above highlight some points of the characteristics of optimization problems which are considered difficult. Initially, it is noticed that despite the use of different taxonomies, researchers address the function characterization problem by creating several categories to describe a particular aspect of a function. Moreover, a lot of the categories are quite similar. For example, (MERSMANN *et al.*, 2011) defines a category called Plateau, whereas (MALAN; ENGELBRECHT, 2013) uses the term Neutrality. It is valid to add that although (ABELL *et al.*, 2013) do not create a specific category related to the plateau of functions, through metrics in the third the category, namely Random Solutions, it would be feasible to characterize such aspects of a function.

2.3.2.4 Metrics for The Characterization of Optimization Problems

With the different categories of characteristics established it is possible to extract suitable metrics for the characterization of optimization problems. As (BISCHL *et al.*, 2012) state, efforts to develop metrics to characterize optimization problems date back to (JONES; FORREST, 1995), however, a single metric is unable to capture all the particularities of the problem structure. Thus, they comment that a feature-set approach appears to be more pragmatic in that it is up to the researcher to select the most useful metrics through experimentation.

Table 2.1 summarizes some of the characterization metrics. It should be emphasized that the list is by no means exhaustive, serving as an illustration of recurrent metrics in Exploratory Landscape Analysis (ELA) and their relation to characteristics of continuous functions. For a more exhaustive list of characteristics, the interested reader is referred to (KERSCHKE, 2017), where the author implemented a tool using the R programming language, containing approximately 300 metrics of characterization of optimization problems. The actual metrics adopted in the present work, as well as their definitions, are included in Appendix C.

TABLE 2.1 – Definition of ELA Metrics

Metrics for Function Characterization			
Metric	Definition	Characteristic	Reference
Num. Variables	$x_1, x_2, ..., x_n$	Dimensionality	I*
Fitness Distance Correlation	$\frac{Cov(x_1,x_2)}{\sigma_{x1}\sigma_{x2}}$	Evolution Cap., Disinformation	I*
Num. Peaks	$p_1, p_2, ..., p_n$	Modality	I*
Assimetry	$E[(\frac{x-\mu}{\sigma})^3]$	Symmetry	I*
Kurtosis	$E[(\frac{x-\mu}{\sigma})^4]$	Neutrality, Ruggedness	I*
Length Scale	$\frac{\|f(x_1)-f(x_2)\|}{\|\|x_1-x_2\|\|}$	Neutrality, Ruggedness	I*
Neg. Slope Coefficient	$\sum min(p_i, 0)$	Evolution Cap., Disinformation	I*
Dispersion	$\frac{1}{(\alpha n)(\alpha n-1)} \sum \sum \|\|x_i - x_j\|\|$	Evolution Cap., Disinformation	I*
Information Content	$-\sum p \log_6(p)$	Evolution Cap., Disinformation	I*
Partial Information Content	$\frac{\mu}{n}$	Modality	I*
Information Stability	$min(\epsilon, S_i(\epsilon))$	Ruggedness, Neutrality	I*
Coef. Linear Pearson	$R^2Linear$	Meta-Model*	II*
Coef. Linear Model	$min(\beta_1, \beta_2, ..., \beta_n)$	Meta-Model*	II*
Coef. Quadratic Pearson	$R^2Quadratic$	Meta-Model*	II*
Coef. Quadratic Model	$min(\alpha_1, \alpha_2, ..., \alpha_n)$	Meta-Model*	II*
Prob. Convexity	$p(x_n) < f((1-\lambda)x_1 + \lambda x_2)$	Convexity*	II*
Dev. Convexity	μ_c	Convexity*	II*
Mean Error. Lin	μ_{mce}	Level Set*	II*
Fraction Mean Error	$\frac{\mu_{mcelin}}{\mu_{mcequad}}$	Level Set*	II*
Mean Error. Quad	$\mu_{mcequad}$	Level Set*	II*
Mean Error. Mix	μ_{mcemix}	Level Set*	II*
Num. Local Optima	$x_1^*, x_2^*, ..., x_n^*$	Local Search*	II*
G-L Constrast	$\frac{x_g^*}{mean(x_1^*, x_2^*, ..., x_n^*)}$	Local Search*	II*
Prop. Cluster O-P	$\frac{p(x_i)}{p(x_j)}, x_i \in K^*, x_j \in K^w$	Local Search*	II*
Prop. Mean Cluster NO	$\mu_p, \forall x_i \notin K^*$	Local Search*	II*
Dist. Num. Evaluations	$x_0, x_{0.25}, x_{0.5}, x_{0.75}, x_1$	Local Search*	II*
Dist. Norm Gradient	$p.d.f\|\|\nabla g\|\|$	Local Search*	II*
Dist. Max-Min Partial Gradient	$p.d.f\frac{\frac{\partial F}{\partial x_{max}}}{\frac{\partial F}{\partial x_{min}}}$	Local Search*	II*
Dist. Max-Min Eigenvalues	$p.d.f\frac{\lambda_1}{\lambda_2}$	Local Search*	II*

Where I* refers to (MALAN; ENGELBRECHT, 2013) and II* to (MERSMANN *et al.*, 2011).

2.4 Dimensionality Reduction

2.4.1 Introduction

Dimensionality Reduction (DR) techniques have been receiving attraction as a result of the increasing need to handle the huge amount of data (GRACIA *et al.*, 2014). By reducing the number of dimensions while at the same time preserving the core information of a dataset, latent variables and hidden patterns from a higher dimensional data set may suddenly become visible. In this sense, this newly expressed information is useful for further exploration such as in the case of the discovery of groups of similar data (clusters) or anomalous points.

The literature on DR techniques is nothing but short as many algorithms have been proposed by researchers during the years. In (CUNNINGHAM; GHAHRAMANI, 2015), the authors perform a thorough literature review in linear DR techniques, where PCA is one of its main techniques. They compare many different techniques among which the aforementioned PCA, Factor Analysis, Multidimensional Scaling. Also, they suggest a framework in which the DR problem formulation may be viewed as an optimization problem, where the idea is to optimize some function of features (distance between pair of points, for example), subject to the matrix manifold, which captures the overall structure of the dataset.

However, due to no inherent nonlinear nature of many natural phenomena, linear DR techniques may not be suitable for rightfully preserving the structure of such datasets. Therefore in (LEE; VERLEYSEN, 2007), the authors not only present the comparison of different nonlinear techniques but establish important principles on which a DR should function. For example, they cite the Curse of Dimensionality and the empty space phenomena. They also compare different strategies for information preservation of DR techniques such as spatial distance, graph distance, and topology preservation. All of which related to different algorithms that may embed (project) differently the same dataset.

2.4.2 The Concept of Manifold Learning

(LI *et al.*, 2019) state the existence of different approaches for Dimensionality Reduction. Some techniques are based on finding a lower-dimensional compact representation by leaning on statistics approaches. Other DR approaches focus on the preservation of the geometric structure of the original data, thus making it more suitable for visualization purposes.

The present work adopts DR via Manifold Learning. (ZHENG; XUE, 2009) provide a loose definition of manifolds, which are topological spaces that locally resemble a Euclidean

space. Therefore, (PLESS; SOUVENIR, 2009) provides one formalization of the Manifold Learning problem:

Given an input set X, which is a finite subset of $\Re^D$, for some dimension D, learn a parametrization that produces a mapping function $f : X \mapsto \Re^d$, which preserves some properties of the structure of X.

The next subsections provide the definitions of the PCA and t-SNE techniques, both chosen to be further studied in the present work.

2.4.3 Principal Component Analysis

Given a dataset on n dimensions, PCA aims to find a linear subspace of dimension d lower than n such that data points lie mainly on a subspace called 'principal components'. They have the characteristics of being an orthogonal and linear transformation of the original data points (GHODSI, 2006).

Besides, (JOLLIFFE; CADIMA, 2016) describe the basic method for employing the PCA technique: Given a NxP data matrix X, seek a linear combination of the columns of matrix X with maximum variance, in other terms, $\sum_{j=1}^{p} a_j x_j = Xa$, where a is a vector of constants $a_1\, a_2, ...a_p$. Thus, the variance of the linear combination may be written in the form $var(Xa) = a'Sa$, with S the sample data covariance matrix. Therefore, in optimization terms, obtain the a $p-$dimensional vector which maximizes $a'Sa$. The authors state that for the problem to be well defined, a constraint must be imposed, such as $a'a = 1$, for example. The model can be written in a Lagrangean approach $a'Sa$-$\lambda(a'a - 1)$, where λ is the Lagrange multiplier and promptly differentiated thus producing $Sa - \lambda = 0 \Longleftrightarrow Sa = \lambda a$.

2.4.4 t-Stochastic Neighborhood Embedding

The t-SNE algorithm (MAATEN; HINTON, 2008) has been an increasingly popular DR technique in the field of Machine Learning. Different than the PCA, the t-SNE deals with nonlinear reduction (transformation) of the original data. As defined in their original article, the t-SNE transforms euclidean distances between data points into conditional probabilities. This procedure is applied in both higher and lower-dimensional space. Then, an optimization routine is performed to minimize the conditional probabilities produced by the higher and lower dimensional data.

In the upper dimensional space, the conditional probability is supported by a Gaussian

distribution and σ_i its the standard deviation centered in point i:

$$p_{ji} = \frac{exp(-||x_i - x_j||/2\sigma_i)}{\sum_{k \neq i} exp(-||x_i - x_k||/2\sigma_i)} \tag{2.1}$$

Whereas in the lower dimensional space, the conditional probability is supported by the Student- t distribution and is defined as follows:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}} \tag{2.2}$$

Then, the Kullback-Leibler Divergence is defined as a cost function:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{ji} log(\frac{p_{ji}}{q_{ji}}) \tag{2.3}$$

Finally, the gradient of the cost function is calculated as follows:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1} \tag{2.4}$$

The authors have provided a simple algorithm employing the t-SNE algorithm:

Data: $\{X = x_1, x_2, ..., x_n\}$;
Cost Function Parameters: Perplexity $Perp$;
Optimization Parameters: Number of Iter. T, Learning Rate η, Momentum α ;
Result: Low-Dimensional Data Representation $\gamma^T = \{ y_1, y_2, ..., y_n \}$
Begin;
  Compute pairwise affinities $\rho_{j|i}$ with perplexity $Perp$;
  Set $\rho_{ij} = \frac{\rho_{j|i} + \rho_{i|j}}{2n}$;
  Sample initial solution $\gamma^0 = \{y_1, y_2, ..., y_n \}$ from $N(0, 10^{-4}I)$;
for *t = 1 to T do* **do**
  Compute low-dimensional affinities q_{ij};
  Compute gradient $\frac{\partial C}{\partial \gamma}$;
  Set $\gamma^t = \gamma^{t-1} + \eta \frac{\partial C}{\partial \gamma} + \alpha(t)(\gamma^{t-1} - \gamma^{t-2})$;
end
End;

Algorithm 1: The t-SNE DR procedure. Adapted From (MAATEN; HINTON, 2008)

The algorithm provided in (MAATEN; HINTON, 2008) indicates, that for the use of the t-SNE, some parameters must be provided. First, the Perplexity P_i is input from the user. The Perplexity reflects the effective size of the neighborhoods in which the t-SNE applies its similarity procedure. Other tuning parameters are related to the optimization

routine, such as the number of iterations T, learning rate η and momentum $\alpha(t)$ (MAATEN; HINTON, 2008).

2.4.5 Quality Preservation: The Co-Ranking Matrix

2.4.5.1 The Concept of Rank

To assess the quality of the projected, the present dissertation makes use of the Co-Ranking matrix concept (LEE; VERLEYSEN, 2009). Figure 2.2 gives an intuition of the concept of the rank between two data points.

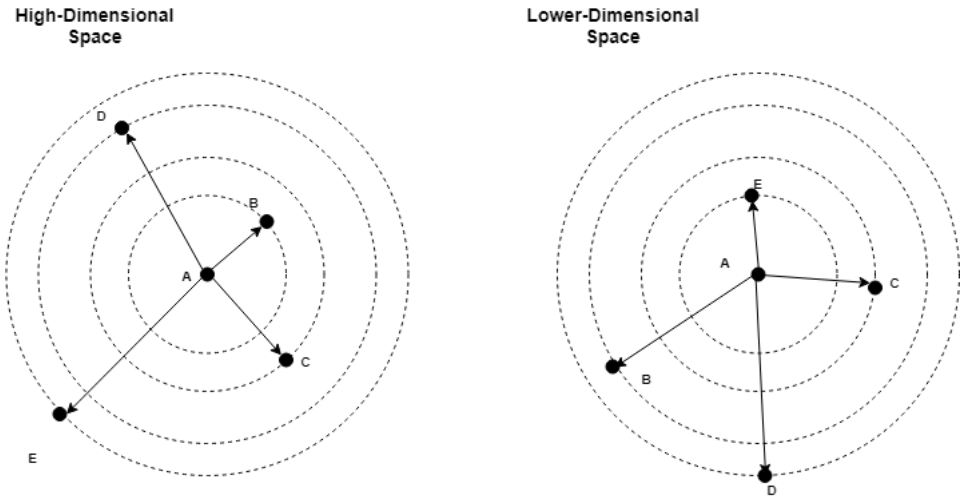


FIGURE 2.2 – Concept of ranking between points

Assume a given set of data points $\{A, B, C, D, E\}$ in space $\Re^D$. The rank of a point j in respect to i is simply the number of points inside the neighborhood defined by the ball Dij. In Figure 2.2, in the higher dimensional space, the ranks of points $\{B, C, D, E\}$ in respect to point A are $\{1, 2, 3, 4\}$, respectively. However, as previously mentioned in the subsection of Dimensionality Reduction, representing a manifold in a lower-dimensional may lead to some distortion in the embedded data. Let's also assume that after the manifold learning algorithm was employed to a dimension $\Re^d$, with $d < D$, the distance of points in the set are depicted in the right side of Figure 2.2. Therefore, the new ranks of point $\{B, C, D, E\}$ are $\{3, 2, 4, 1\}$. It is noticeable that an increase of rank, such as the case of point B, when the ranking increased from 1 to 3, points got farther than in the true, higher-dimensional space. Conversely, point E was the most distant from point A and became the closest after the DR procedure.

The Co-Ranking framework addresses this specific rank swap. If the data has been successfully embedded in a lower-dimensional space, all distances between points (hence their neighborhoods and ranks) must also have been preserved.

2.4.5.2 Mathematical Definition of the Co-Ranking Matrix

The rank of x_{ij} with respect to x_i in the high-dimensional space is written as $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij}\, or\, (\delta_{ik} = \delta_{ij}\, and\, k < j)\}|$, where $|.|$ denotes the set cardinality. Similarly, the rank of x_j with respect to x_i in the lower-dimensional space is $r_{ij} = |\{k : d_{ik} < d_{ij}\, or\, (d_{ik} = d_{ij}\, and\, k < j)\}|$. Hence, reflexive ranks are set to zero and ranks are unique.

Then, the authors continue the definition of the co-ranking matrix such as:

$\mathbf{Q} = [q_{kl}]_{1 \leq k,l \leq N-1}$ with $q_{kl} = |\{(i,j) : \rho_{ij} = k\, and\, r_{ij} = l\}|$

From such above definition, one may notice that if the embedded dataset was perfect, all pair of points i and j should preserve their rank and therefore the Co-Ranking matrix should consist of only diagonal entries. Once the framework established, suitable metrics to assess the quality of the embedding are inferred. In (KRAEMER *et al.*, 2018) the authors present several of these metrics.

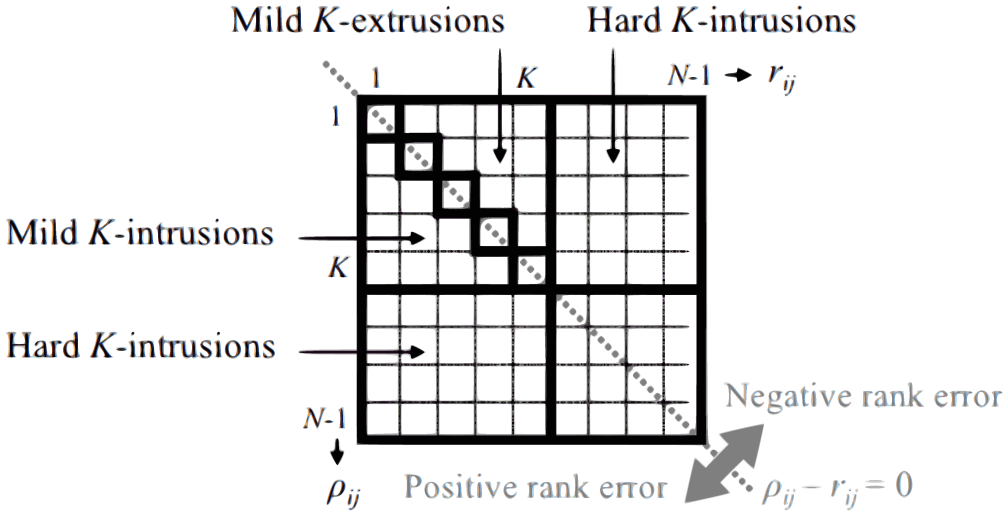


FIGURE 2.3 – The concept of the Co-Ranking Matrix. From (LEE; VERLEYSEN, 2009)

2.4.5.3 Quality Criteria from the Co-Ranking Framework

As previously stated, a perfect embedding of data produces a Co-Ranking matrix with only diagonal entries. Therefore, suitable metrics can be extracted for quality assessment. The simplest of these is to count the number of rank swaps according to a specified neighborhood of size k:

$$Q_{NX}(k) = \frac{1}{kn} \sum_{i=1}^{k} \sum_{j=1}^{k} q_{ij} \tag{2.5}$$

However, (CHEN; BUJA, 2009) raised the issue that many of the points inside a $k-ary$ neighborhood may have been due to chance. Therefore, a new metric is established to remove the random factor. The criterion is called Local Continuity Meta Criterion and is defined as follows:

$$LCMC(k) = Q_{NX}(k) - \frac{k}{n-1} \tag{2.6}$$

Finally, because the $LCMC(k)$ possesses a defined upper bound (LEE; VERLEYSEN, 2007), a final metric may be defined, where the $LCMC(k)$ is standardized in the range from 0 to 1:

$$R_{NX}(k) = \frac{(n-1)Q_{NX}(k) - k}{n-1-k} \tag{2.7}$$

The RNx quality shall guide an optimization procedure where only the metrics which best aid in the preservation of the lower-dimensional manifold will be selected.

2.5 The Algorithm Selection Problem

Since the 70's the Algorithm Selection Problem (ASP) was already a concern from part of researchers. (RICE, 1976) proposes the model in figure 2.4, where he describes the ASP.

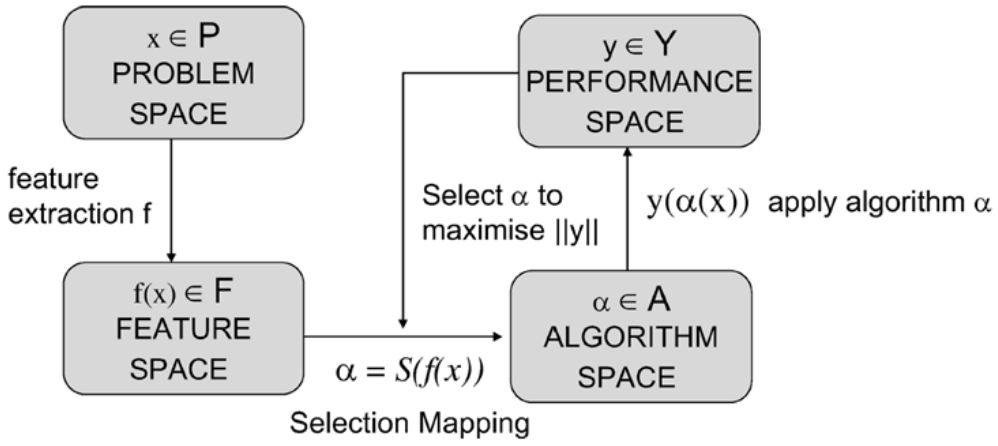


FIGURE 2.4 – The Algorithm Selection Problem. From (RICE, 1976)

- **Problem Space**: The set which contains all problem instances of a determined problem category.

- **Feature Space**: The set which contains all necessary features to characterize the problem instances.
- **Algorithm Space**: This set contains all algorithms suitable for solving the problem in question.
- **Performance Space**: The set which contains all metrics necessary to characterize the ability of a certain algorithm in solving a specific problem.

In the framework proposed in (RICE, 1976), one may notice that the choice of an algorithm is made to maximize the performance of an algorithm for a specific problem. The present work is confined to the box named Feature Space and Selection Mapping in Figure 2.4.

2.6 Related Work

There has been a lot of interest in the research community for the automated selection of algorithms. With more computing power available and Machine Learning techniques, one does not need to resort to trial and error to select an appropriate algorithm for one's problem (LEMKE *et al.*, 2015). Also, with the availability of several different ELA metrics, different models may be built based on different characteristics of a given function.

The field of Fitness Landscape Analysis has been active since early in (JONES; FORREST, 1995), when the author introduced the Fitness Distance Correlation analysis to assess the difficulty of Genetic Algorithms in finding the global optima of functions. Since then many authors have attempted to characterize optimization functions according to some metric. In (LUNACEK; WHITLEY, 2006), the authors introduce the Dispersion Metric as a means for assessing the reasons why the CMA-ES optimization algorithm would perform better in some function instances than others. In (MALAN; ENGELBRECHT, 2009), the authors use the concept of Information Content introduced by (VASSILEV *et al.*, 2000) to characterize the ruggedness of continuous functions. For a more thorough development of the field throughout the years, the interested reader is referred to (MALAN; ENGELBRECHT, 2013) and (KERSCHKE, 2017).

However, as it may have become evident, since continuous functions may possess different characteristics, it may not be reasonable that a single metric can capture all information contained in a given landscape. In (MORGAN; GALLAGHER, 2017), (SUN *et al.*, 2014), (MERSMANN *et al.*, 2010), (MUÑOZ; SMITH-MILES, 2015) and (MUÑOZ; SMITH-MILES, 2017),for example, the authors approach the characterization of functions of the BBOB set of functions in with a "bag" of metrics approach. Contrary to what is proposed in the present work, the selected features in the aforementioned works were chosen qualitatively.

That is, features that are believed to measure some aspect of the landscape (e.g: Modality) were included in the set of metrics.

In the theme of ASP, In (CLUNE, 2013), the author proposes an algorithm selection framework in the field of structural engineering. The author chooses 8 algorithms for both local and global optimization and tests them in a variety of representative designs such as trussed arches, girders and suspension only to name a few. Then, he studies two machine learning tasks, namely classification and regression to map the relation between problem design and algorithm performance. In (LANG *et al.*, 2015), the authors used an automatic model selection for high-dimensional survival analysis data. Relying on 4 regressions models, and with the aid of automatic configuration of algorithms, they were able to produce model selection via techniques of automatic configuration of algorithms.

Within the theme of the benchmark of continuous functions, there have also been closely related works. In (CUI *et al.*, 2016), the authors used the set of functions from IEEE CEC 2013/2014 to build a recommender system for meta-models. The IEEE CEC 2013/2014 benchmark shares some of the functions with the BBOB 09. However, in (CUI *et al.*, 2016), the authors were mainly concerned in choosing the fittest meta-model for capturing function's overall tendencies (surrogate modeling), whereas the present work is concerned with choosing the ELA metrics (features) which best represent the function categories in a lower-dimensional space. Similar, (MERSMANN *et al.*, 2010) do employ feature selection of ELA metrics, along with the BBOB 09 set of functions algorithm selection. However, in their work, they use the validation training for feature selection in a feed-forward scheme.

Finally, researchers have also been interested in the combination of concepts of ELA metrics with Dimensionality Reduction techniques. In (MORGAN; GALLAGHER, 2017), the authors propose a new family of ELA metrics called Length Scale. They compare their newly proposed metrics to a set of traditional metrics. The t-SNE technique was employed to depict in a 2D scatter plot how the different sets of ELA metrics were able to distinguish between the functional categories of the BBOB09 set. The proposed Length Scale metrics did indeed manage to capture the intrinsic characteristics of the data. Similarly, in (MUÑOZ; SMITH-MILES, 2015), the authors use a bag of ELA metrics to characterize the BBOB09 function, however, they apply the PCA technique for the task embedding the higher dimensional data. In (MUÑOZ; SMITH-MILES, 2017), the authors also employ PCA and a wider set of ELA metrics for the characterization of the BBOB09 functions. However, their interest was in identifying the regions of the embedded space where a given optimization algorithm would perform best. Although they have been able to suggest the best algorithms for given regions, the authors have found "anomalous" regions, that is, the suggested algorithm didn't behave as prescribed by their model. In none of the cited works the authors tried to measure the distortion of the embedded

manifold, that is, what is being depicted in the 2D space doesn't necessarily reflect the structure in the higher dimensional data.

In this sense, the present work contributes to the research in the field of ELA and ASP. By employing a specific framework for assessing the quality of the embedded data, it is hypothesized that an awareness of distortions in the 2D lower-dimensional space, future analysis will be made with caution. Also, the procedure described for feature selection may serve as inspiration for future ELA metrics selection. Finally, the set of selected features may also reveal interesting, hidden characteristics of the ELA metrics or validate what has been proposed so far in the literature. In Chapter 3: Materials and Methods the experimental design of the present work is described.

3 Materials and Methods

3.1 Introduction

In this chapter, the main steps of the selection of Exploratory Landscape Analysis metrics are described. At first, there is a need for carefully selecting a suitable set of continuous functions that shall be characterized by sets of ELA metrics. Then, after the calculation of ELA metrics proposed by different authors (MORGAN; GALLAGHER, 2017) (SUN *et al.*, 2014), a data pre-processing takes place, that is, the dataset is cleaned by the removal of redundant information or missing values. In a third step, the feature selection procedure is performed. From a heterogeneous superset of ELA metrics, an optimization algorithm is run to select the few metrics that preserve the most information about the structure of the data. The idea is to compare the selected set of features from a Dimensionality Reduction procedure with the set o features already established in the literature.

Initially, the sets of ELA metrics are depicted in density plots. This first assessment provides insight on how each ELA metric distinguishes functions in the test set by their categories. In a second step, DR techniques project (embed) each set of ELA metric, so data points from a higher-dimensional space may be displayed in a 2D lower-dimensional space. This analysis is useful, since it shows visually the similarity between the categories of functions, as well as provide information on the possible clusters and hidden structures in the data. Finally, the fidelity of the embedded data in the 2D space shall be assessed by an information quality metric from the Co-Ranking framework (LEE; VERLEYSEN, 2009). The selected quality metric provides a quantitative assessment of the degree of distortion of the embedded data incurred by the Dimensionality Reduction procedure. Besides, the Co-Ranking Framework provides information about the performance of selected DR techniques, as well as their limitations in preserving local and global structures of the higher dimensional dataset.

3.2 Selection of Exploratory Landscape Analysis metrics: A Conceptual Framework

In the present work, three main steps are proposed for achieving the objective of selecting ELA metrics for the characterization of continuous functions. These steps are depicted in Figure 3.1.

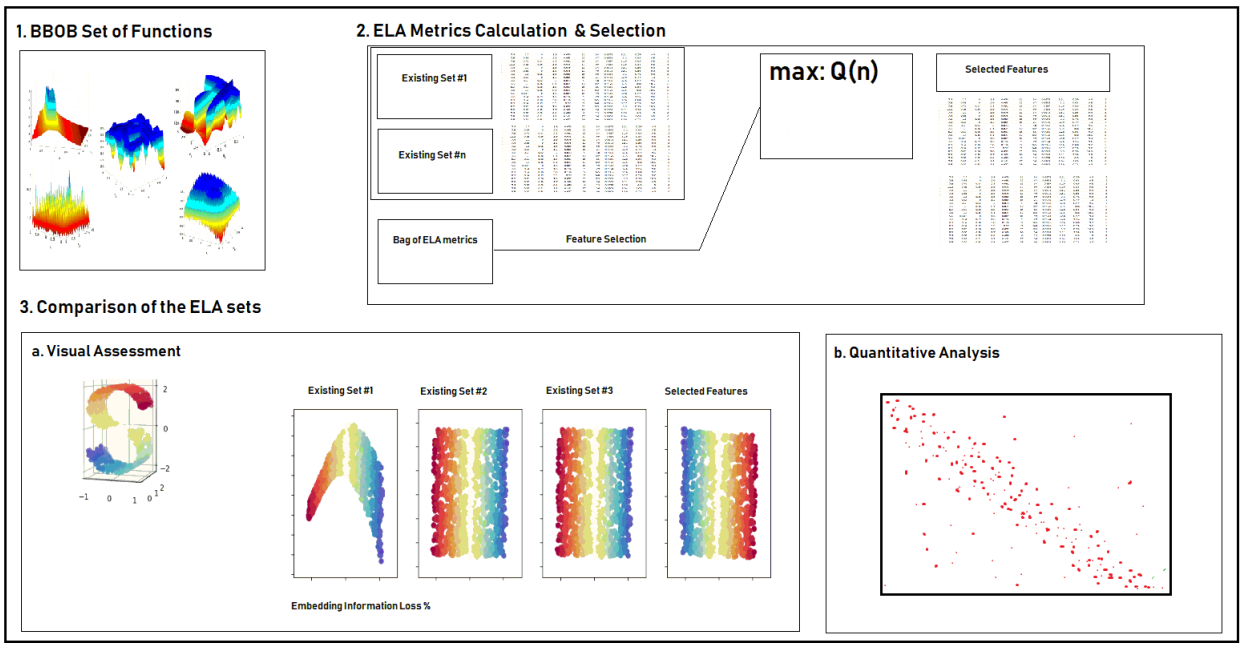


FIGURE 3.1 – The Conceptual Framework of the Present Work

A high level description of each stage follows:

I. **Selection of Test Functions**: In this initial stage the focus lies on the selection of a suitable set of test functions. As mentioned throughout the chapters, the BBOB09 benchmark (HANSEN *et al.*, 2009) shall be used. Therefore, the type of functions, number of instances and number of dimensions are the configurations that shall be defined so a test set may be generated.

II. **ELA Metrics Calculation and Selection**: In the second stage, a set of ELA metrics proposed by different authors are calculated. These shall be used to compare to the sets of metrics generated by the present work. Because many of these metrics are based on randomly generated samples, some replications are performed to capture uncertainty around each metric. In this same stage, feature selection shall take place. From a wider set of many ELA metrics, a feature selection procedure is performed. The objective is to select the ELA metrics which best preserve the structure of the higher dimensional feature set. The Co-Ranking framework (LEE;

VERLEYSEN, 2009) alongside with a DR technique was adopted. In this manner, the feature selection is seen as an optimization problem where the objective is to maximize a quality information criterion metric given a specified number of features.

III. **Comparison of the ELA metric sets**: Finally, the final stage consists of the comparison of the sets of ELA metrics. This final stage is divided into two steps. First, the comparison is performed visually. Density plots will depict how each ELA metric distinguishes function categories. Then, DR techniques shall be employed and the embedded data depicted in a 2D plot for comparison. In a final step, the Co-Ranking matrices formed by each embedded and original data provides the means for quantitative assessment of information preservation in the lower dimensional space.

3.3 Justification for the Experimental Setup

As already established, the procedure of feature selection entails the use of a wide different number of concepts and techniques. In the present section, the main justifications and hypothesis are retaken to guide the explanation of the experimental design.

3.3.1 The Need of Feature Selection in ELA

Figures 3.2 and 3.3 depict some conceptual functions and their underlying topologies. Not only these functions show the different characteristics that may render function optimization a difficult task, but real-world functions could be any combination of these structures, making the optimization problem even harder.

Herewith, it is not expected that a single feature (ELA metric) will be able to describe all different nuances and characteristics of optimization functions. The research community has shown that approaching the function characterization problem with a different set of features has produced positive results (MORGAN; GALLAGHER, 2017)(GONÇALVES, 2018)(MUÑOZ; SMITH-MILES, 2015).

Besides, a proper characterization of continuous functions may render it easier for the optimization community to understand which algorithms are most suitable to a specific problem. For instance, one may interpret this as related to the No Free Lunch theorem (WOLPERT; MACREADY, 1997), which states that a gain of performance of an algorithm solving a given class of problems will be compensated by a loss in another class. In this way, most algorithms tend to perform similarly on average. Therefore, one might assume that some algorithms are more suitable to a particular class of problems, further

strengthening the justification for better function characterization and its utility for an adequate algorithm selection process.

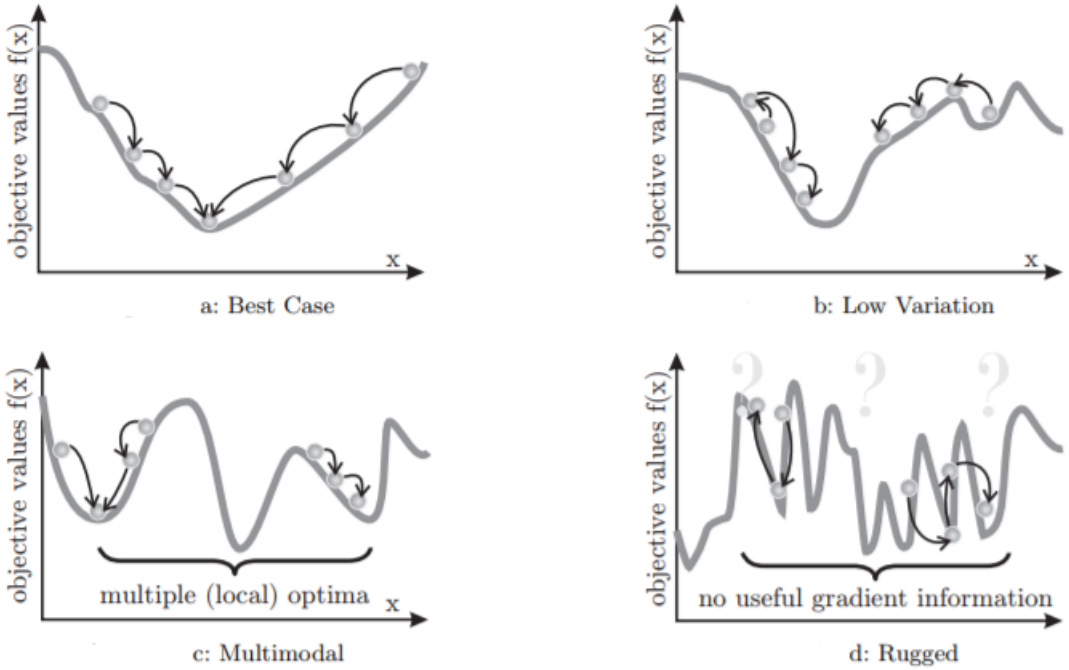


FIGURE 3.2 – Function Characteristics - A. From (WEISE, 2009)

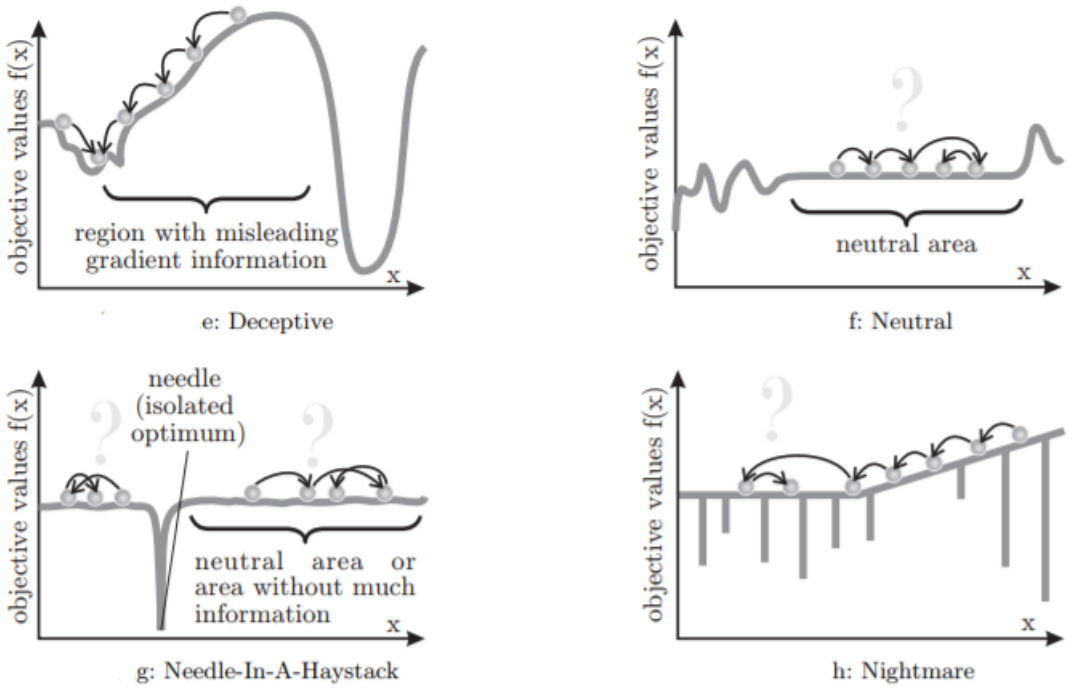


FIGURE 3.3 – Function Characteristics - B. From (WEISE, 2009)

3.3.2 Hypothesis: Feature Selection through Dimensionality Reduction

In this subsection, the hypothesis of the present work is presented. It is formulated as follows:

DR Techniques are useful for selecting ELA metrics for characterization of continuous functions.

Therefore, the hypothesis itself is based on underlying questions that shall guide the experimental design. One first question is the set of test functions. Though it has been already established the BBOB09 functions are to be used (HANSEN *et al.*, 2009), many authors use different configurations of this benchmark to create their set of test functions. The selected configurations will serve as support to the elaboration of the present work test set. A second question is related to which ELA metrics are to be included in a superset of metrics where feature selection takes place. The research literature on ELA metrics has produced a variety of metrics that covers different characteristics of continuous functions. Thirdly, though it has been established DR techniques will support the selection procedure, topics such as the model formulation and the selection of a suitable optimization algorithm must be addressed. Finally, a final question is how the proposed sets of ELA metrics compare to sets already established in the literature to a criterion that will also be defined.

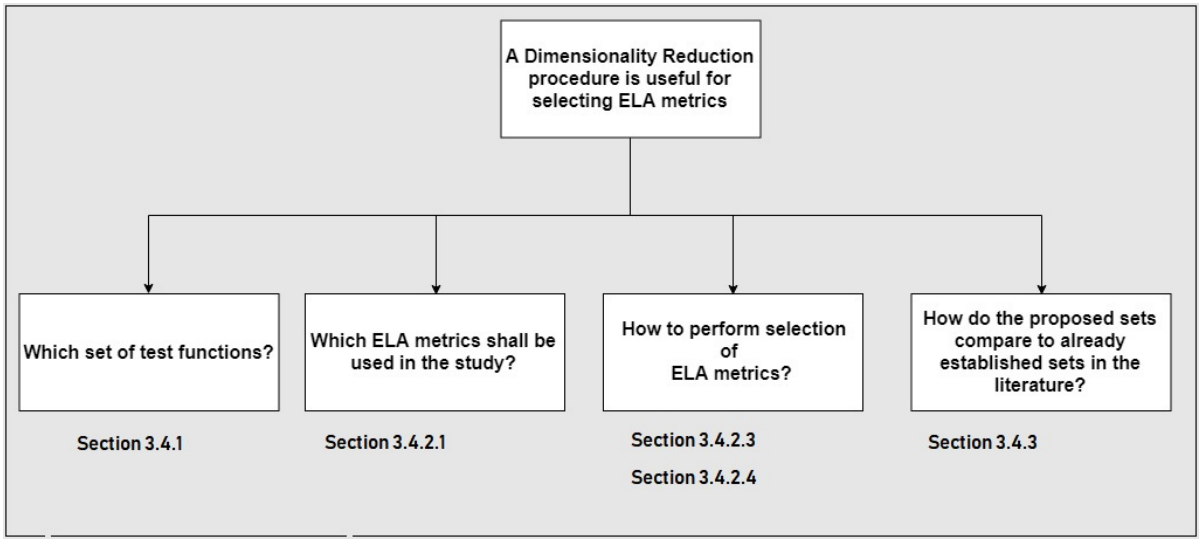


FIGURE 3.4 – Overall Framework Adopted in the Present work.

In Section 3.4 Experimental Design, the experimental setup will be detailed.

3.4 Experimental Design

This section describes the computational experiments that were performed in this work. The R programming language is adopted to perform all calculations. It possesses a strong Statistics, Optimization and Machine Learning community providing documentation and tutorials as well as packages with built-in statistical models. Table 3.1 resumes all packages employed in each of the steps for proceeding with the task of ELA metrics selection.

TABLE 3.1 – Employed R Packages

Package Name	**Use**
smoof: Single and Multi-Objective test Functions.	To create a set of BBOB functions with different instances and number of dimensions.
flacco: Feature-Based Landscape Analysis of Continuous and Constrained Optimization Problems.	It provides the capability of calculating Exploratory Landscape Analysis metrics from black box optimization functions.
dimRed: Dimensionality Reduction	Implementation of Dimensionality Reduction techniques as well as Co-Ranking metrics.
GA	Implementation of the Genetic Algorithm optimization method.

3.4.1 Generation of the Set of Test Functions

The R package smoof (BOSSEK, 2017) is used to generate different instances of the benchmark BBOB set of functions (HANSEN *et al.*, 2009). The smoof package (a.k.a Single and Multi-Objective Optimization Test Functions) implements a collection of known test functions in the optimization research community.

Specifically, there has been a wide adoption of the BBOB 09 set of functions in the research community since it includes a set of functions with rather distinct, intrinsic characteristics. Many of these works were related to simple Exploratory Landscape Analysis of the set of functions, whereas others went further and explored how ELA metrics together with Machine Learning models could aid in the selection of the fittest algorithm (Algorithm Selection Problem) for a given function issued from the BBOB 09 test set. Table 3.2 resumes the configurations of the BBOB09 test functions adopted by different

researchers.

TABLE 3.2 – Adopted BBOB Test Sets in the literature.

Function Type	**Dimensions**	**Instance**	**Sample Size**	**Purpose**	**Reference**
1-24	2, 5, 10, 20	1-30	$1000D^2$	Exploratory Landscape Analysis	(MORGAN; GALLAGHER, 2017)
1-24	2	1-30	-	Exploratory Landscape Analysis	(MUÑOZ; SMITH-MILES, 2015)
1-24	2, 5, 10, 20	1-15	1000D	Exploratory Landscape Analysis	(MUÑOZ; SMITH-MILES, 2017)
1-24	10	1-15	500	Algorithm Selection	(BISCHL *et al.*, 2012)
1-24	2, 3, 5, 10, 20	-	1000D	Algorithm Selection	(MUÑOZ *et al.*, 2012)
1-24	2, 3, 5, 10, 20, 40	1-5	50D	Algorithm Selection	(KERSCHKE; TRAUTMANN, 2019)
1-24	2, 3, 5, 10	1-20	100D	Algorithm Selection	(GONÇALVES, 2018)

It is interesting to note that all researchers adopt the full set of 24 generating functions, but use a different range of dimensions and instances. A particular instance of a function is a rotation or translation in the search space, therefore changing the function upper and lower bounds, as well as the global optimum and fitness values. Also, each author adopted a different strategy concerning sample size. Because many of the ELA metrics such as mean or standard deviation are calculated in sampled values, a strategy for sample size definition is also adopted differently. For example, (MORGAN; GALLAGHER, 2017) calculates features according to 1000*D, where D is the specific dimension of a function.

To avoid the set of test functions of growing too much in size, and using the previous similar work as a reference, the present work adopts the following set of test functions:

- Function ID: 1 - 24
- Dimensions: 2, 5, 10, 20
- Instance ID: 1, 2, 3, 4, 5
- Sample Size: 100*D

Wish the established configuration a total of 480 function instances (24 x 4 x 5) were generated to be further studied. Figure 3.5 depicts instances of the BBOB09 functions 20-24. It is clear the differences between their landscapes.

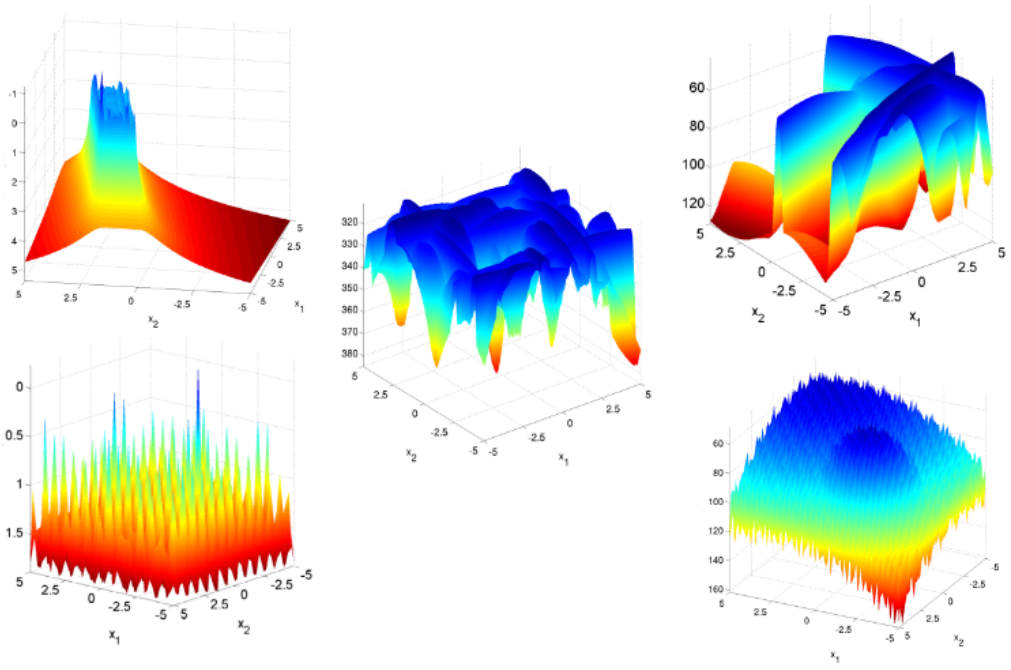


FIGURE 3.5 – Examples of 2D functions in the BBOB set

3.4.2 Feature Set Calculation and Feature Selection

3.4.2.1 Implementation of Established ELA Metrics

Once the set of test functions is defined, the next step is to extract their underlying structure by the calculation of ELA metrics. In the present work, the selected ELA metrics shall be compared to sets of metrics established in the literature. Table 3.3 resumes which metrics are included in each set, whereas their definition is given in Chapter 2: Literature Review, Appendix B and C.

All the ELA metrics defined were implemented in the R script. As noted in subsection 3.4.1: Generation of the Set of Test Functions, the adopted sample size is 100*D. That is, given a function instance with the number of dimensions 20, a sample of 2000 points following a Latin Hypercube Sampling scheme (MCKAY, 1992) is generated.

TABLE 3.3 – Established Set of ELA Metrics - A

Set ID	**ELA Metrics**	**Reference**
#1	Fitness Distance Correlation (Global) Fitness Distance Correlation (Local) Correlation Length Median Dispersion 0.25 Information Content Partial Information Content Information Stability	(MORGAN; GALLAGHER, 2017)
#2	Mean Length Scale Median Length Scale Mode Length Scale Maximum Length Scale Minimum Length Scale Kernel Bandwidth Length Scale Entropy Length Scale	(MORGAN; GALLAGHER, 2017)
#3	Number of Dimensions Fitness Distance Correlation (Global) Number of Peaks Skewness Kurtosis Median Length Scale Information Content Partial Information Content Information Stability	(SUN *et al.*, 2014)

TABLE 3.3 – Established Set of ELA Metrics - B

#4	MMCE LDA 10 MMCE LDA 25 MMCE LDA 50 MMCE QDA 10 MMCE QDA 25 MMCE QDA 50 MMCE MDA 10 MMCE MDA 25 MMCE MDA 50 LDA_QDA 10 LDA_QDA 25 LDA_QDA 50 LDA_MDA 10 LDA_MDA 25 LDA_MDA 50 QDA_MDA 10 QDA_MDA 25 QDA_MDA 50	(MERSMANN *et al.*, 2010)

Overall, the four sets of ELA metrics adopted present rather distinct characteristics except for sets M and SU, which motivates their selection in the present study. Also, these specific sets have been tested in the BBOB09 benchmark set of functions. Therefore, they will serve as the basis for comparison to the proposed sets of ELA metrics and validate any eventual findings from this work.

Sets M and SU are the most similar since they are both comprised of "bag of metrics". Set M adopts ELA metrics related to the Ruggedness and Neutrality of a function (FDC Global, FDC Local, and Correlation Length), to the Information of a function (Information Content, Partial Information Content, and Entropy) and the function Dispersion (Median Dispersion 25%). Set LS is a set based entirely on the statistics of Length Scale metric, which measures the function Ruggedness and Neutrality. Set SU includes metrics based on Ruggedness and Neutrality (FDC Global, Median Length Scale), Fitness (Y) statistics (Skewness, Kurtosis, Number of Peaks), Information (Information Content, Partial Information Content and Information Stability) as well as a metric that indicates the number of dimensions in the function instance. Finally, set MER includes only ELA metrics related to a function level set.

To calculate the ELA metrics of a continuous function there is a need for sampling data

points. It is believed that a rather sparse sample is needed to capture the different changes in the landscape (characteristics) of a continuous function. In this work, the sampling scheme adopted is the Latin Hypercube Sampling (LHS). It has been adopted in similar works in ELA metrics such as in (KERSCHKE; TRAUTMANN, 2019) and (KERSCHKE, 2017).

In an LHS scheme, as shown in Figure 3.6, a grid is formed by discretization of each dimension, thus forming a window. Only a single point is allowed to be taken from each window. In this sense, a rather sparse and uniform sample is taken from a high dimensional space. The number of sample data points (observations) in the sample is defined as 100*D, where D is the dimensionality of a given function.

Moreover, because many features rely on the statistics of the sample (e.g Minimum Dispersion 5%), the procedure of sampling (calculation of ELA metrics) is replicated 100 times to account for uncertainty. The final values of each ELA metric in a given function instance (data point) is their median value calculated from the replication pool.

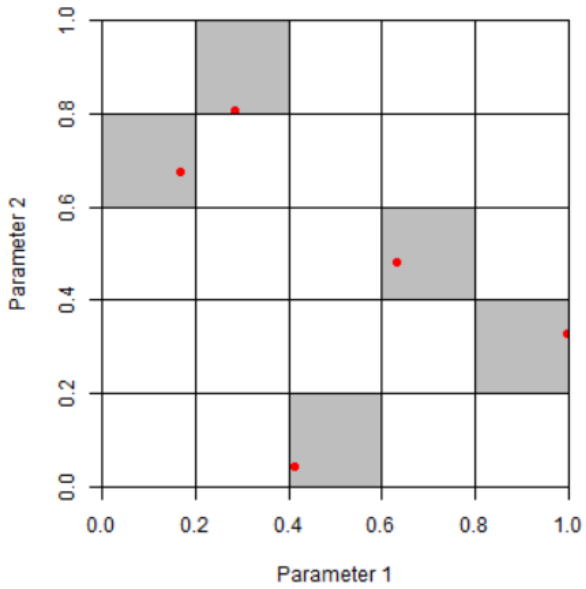


FIGURE 3.6 – Latin Hyper Cube sampling scheme.

3.4.2.2 Definition of the Super Set of ELA metrics

After the calculation of four different sets of established ELA metrics, the following step consists of performing ELA metrics (feature) selection with a Dimensionality Reduction approach. The general idea lies in, from a superset of different ELA metrics, select only the metrics that preserve the most the data set original structure (manifold) in a lower-dimensional space.

The R package flacco (Feature-Based Landscape Analysis of Continuous and Constrained Optimization Problems) (KERSCHKE, 2017), developed by Pascal Kerschke from

TABLE 3.4 – Categories of ELA Metrics in flacco package.

ELA Category	Flacco Identification	Properties of Metrics
Convexity	ela_conv	Measures the probability of convexity in randomly selected points.
Curvature	ela_curv	Measures statistical properties of the estimated gradient in the sample.
Distribution	ela_distr	Measures statistical properties of the objective function.
Level	ela_level	Measures the misclassification error in separating the sample in two classes.
Local Search	ela_local	Measures local characteristics such as size of basin of attractions and number of peaks.
Meta Model	ela_meta	Defines a linear and quadratic model with the specified sample and assess pearson R-square.
Information Content	ic	Measures the information content of a sample based in the theory of Shannon Entropy.

the University of Muenster was used. The package allows the user to calculate over 300 metrics for the characterization of continuous functions. Specifically, in the present work, all ELA metrics related to the following categories are excluded: PCA, Barrier Trees (HERNANDEZ *et al.*, 2014), Nearest Better Clustering (KERSCHKE *et al.*, 2015) and Cell Mapping (KERSCHKE *et al.*, 2014). In the flacco package, these metrics required extra computing time, thus making it infeasible to provide quick information about the functions. With these restrictions, a total of 106 ELA metrics are calculated. For further theoretical developments see (MERSMANN *et al.*, 2011), (MERSMANN *et al.*, 2010) and (MALAN; ENGELBRECHT, 2013). Table 3.4 specifies the different categories of ELA metrics implemented in the package flacco that shall be adopted. Also, the reader is referred to Appendix B and C for the definition of the ELA metrics in the present work and the to the table including the metrics in the superset where the selection procedure are based on.

3.4.2.3 Definition of the Optimization Model

A metaheuristic algorithm is employed for feature selection. Initially, a population of random initial solutions is generated. These individuals are represented with vectors of length 106 (the size of the superset of ELA metrics), with real values in the range from 0

to 1. If the value in a given position from 1 to 106 is greater than 0.5 than the feature is selected and its associated column shall be present in a reduced set of ELA metrics.

Then, the reduced set of selected metrics is embedded using a DR technique. With the aid of the Co-Ranking framework, a quality criterion metric is calculated to assess the amount of information preserved in the lower dimensional space. As in an optimization scheme, recombination operators modify the initial solutions forming the new population by selecting the solutions which present higher quality information criterion. The evolutionary optimization approach for selecting features has been successfully applied as seen in (YANG; HONAVAR, 1998), (LEARDI *et al.*, 1992) and (PUNCH *et al.*, 1993). Figure 3.7 illustrates the procedure for feature selection.

Moreover, three different strategies will be tested in the optimization procedure. The reasoning is to test whether the optimization run with these strategies will produce useful insights about the sets of selected ELA metrics. The strategies are explained as follows:

I. Unrestricted Size: There isn't any restriction on the size of the individuals during the optimization procedure. Therefore, any set of ELA metrics may be considered.

II. Set Cardinality - 10 Features: In this strategy, the cardinality of the selected set is restricted to size 10. This is the rounded average number of ELA metrics included in the 4 sets (M, LS, SU, and MER) that were selected to be further studied in this work. Therefore, during the optimization procedure, a penalty term is included in the objective function for penalizing the individuals (sets) with cardinality different than 10.

III. Stratified Initial Population: In this strategy, an initial population with defined characteristics is inserted in the first iteration of the optimization algorithm. The initial population consists of at least $Nind$ examples of individuals that produce a set of ELA metrics of cardinality from {3, 4, 5,...,106}. For e.g, with a $Nind = 2$, there will be 2 individuals with 3 values in their vector indexes over 0.5 with the rest below 0.5, 2 with 4 values over 0.5, until 2 individuals with all values over 0.5 The reasoning behind this strategy is to cover a wide range of sets of different sizes. Note: The indexes of the individual which are greater than 0.5 are randomly generated.

Finally, the objective function of the model was defined as the maximization of the criteria mean R_{NX} (minimization of -R_{NX}). As described in Chapter 2 the metric ranges from 0 to 1 and describes the quality of the embedding. The models to be optimized as defined as follows:

Model Strategies 1-3

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & mean(-R_{NX}(x)) \\ \text{subject to} \quad & x \in [0,1], \qquad \forall x_i \in x,\ i = 1, \ldots, 106. \end{aligned}$$

Model Strategy 2

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \frac{mean(-R_{NX}(x))}{2^k} \\ \text{subject to} \quad & k = \sum_{i=1}^{i=106} x_i, \\ & k = 10, \\ & x \in [0,1], \qquad \forall x_i \in x, i = 1, \ldots, 106. \end{aligned}$$

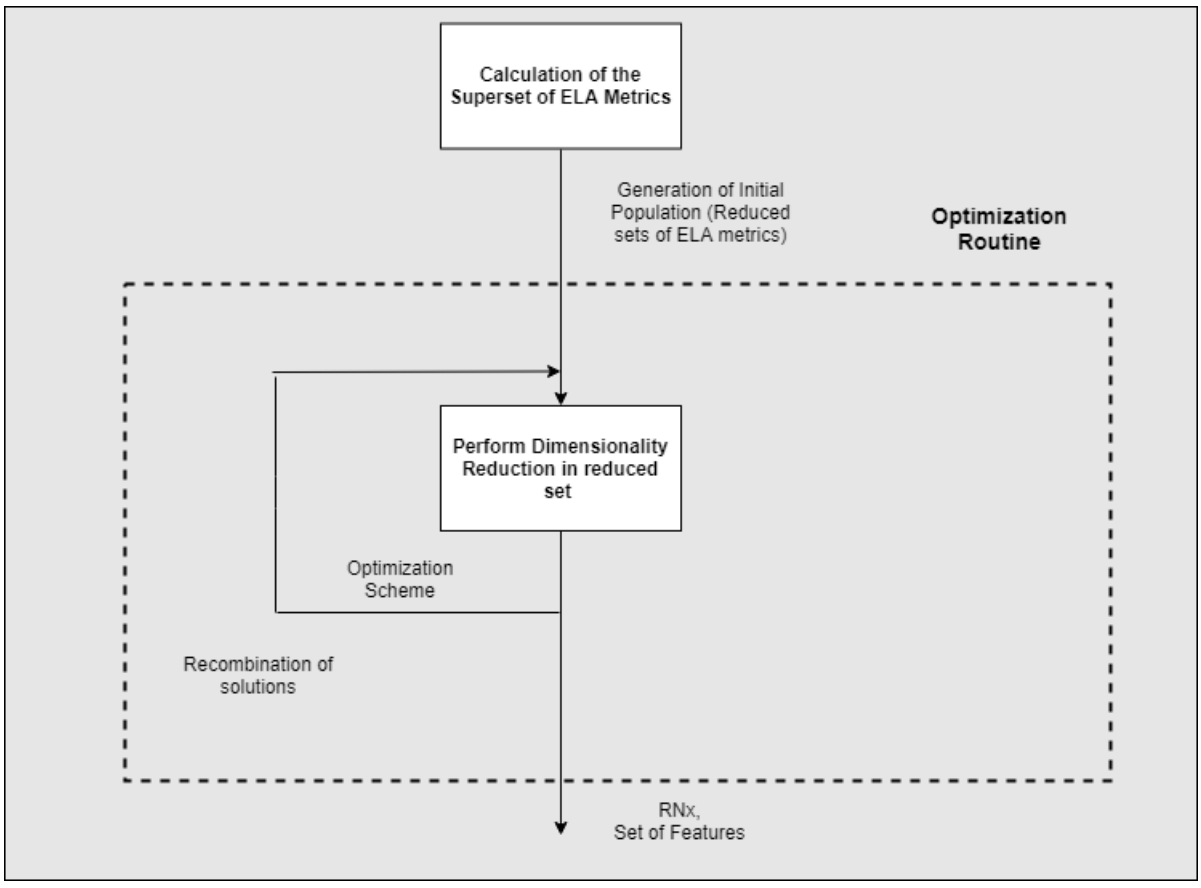


FIGURE 3.7 – Flowchart for Feature Selection.

3.4.2.4 Algorithm Setup

The Genetic Algorithm (GA) (GOLDBERG, 1989) is selected to perform the optimization. The R package mco (MERSMANN *et al.*, 2014) provides a simple implementation of this type of evolutionary algorithm and a function in R language was written to im-

plement the optimization models. As explained in the previous section, an individual is represented as a vector of length 1x106 (total number of features). A feature should be only included in the set of ELA metrics if its associated index in the individual (vector) contains a value greater than 0.5. Then, with the reduced set of selected ELA metrics, a dimensionality reduction algorithm performs an embedding into a lower dimension space. Afterward, the quality of the embedded data is assessed by calculation of the mean R_{Nx} metric.

The GA optimization algorithm is set with the following configurations expressed in the list of items. These are the basic recommendations taken from (MERSMANN *et al.*, 2014). Also, the population size is an important parameter because it ultimately reflects the ability of the individuals in covering the search space. The population size PS will be adopted as 5D, where D is the dimensionality of the problem. For the optimization of the model in Strategy $Nind$ is defined as 5.

- Input Dimension: 560
- Number of Objectives: 1
- Input Lower Bound: 0
- Input Upper Bound: 1
- Population Size: 560
- Number of Generations: 100
- Crossover Probability: 0.70
- Mutation Probability: 0.20
- Selection Type: Tournament Selection of size 2

Dimensionality Reduction is a central part of the present work since only the features that help to preserve most of the information in a lower-dimensional space are selected. Therefore, a proper DR technique is paramount for the success of the selection procedure. In this way, the t-SNE DR (MAATEN; HINTON, 2008) algorithm is selected for producing embedded data during the optimization routine, upon which, the mean RNx metric is be calculated.

Similar to the GA optimization algorithm, the t-SNE also possesses tuning parameters that affect the quality of its produced result. As mentioned in its original article, the perplexity adjusts the effective neighborhood of data point i where the similarity between points is calculated. As mentioned in their original article (MAATEN; HINTON, 2008),

values of perplexity between 5 and 50 have shown to be rather robust. In the present work we adopt the perplexity as P = 30 and $\eta = 0.5$.

The adopted values of the parameters for both Genetic Algorithms and t-SNE were based on related or original articles. However, it is the author belief that a procedure of tuning optimization could indeed improve the results solutions provided by the GA and the embedded data using the t-SNE technique. Due to lack of time, we left the study of the tuning of algorithms and their impact on the quality of the embedded data as a subject for future research.

The algorithm listed in Algorithm 2 further describes the selection procedure.

Result: Set of ELA metrics
Initialize population of size 560;
Set MaxGeneration to 200;
Set Counter to 1;
Set Crossover Probability to 0.70;
Set Mutation Probability to 0.20;
while *While Counter less or equan than MaxGeneration* **do**
  for *Each individual in Population* **do**
    Identify indexes in an individual greater than 0.5;
    Create a set of ELA metrics with features associated with the identified indexes;
    Employ t-SNE for dimensionality reduction;
    Calculate the mean RNx criteria;
    Multiply mean RNx by -1;
  end
  Apply Crossover recombination operator to produce offspring population;
  Apply Mutation operator to offspring population;
  Set current population as the offspring population;
  Increase one unity in Counter;
end

Algorithm 2: Feature Selection based in t-SNE and Genetic Algorithms

3.4.3 Comparison of the ELA sets

The final step in the workflow consists of using Dimensionality Reduction techniques to compare the established sets of ELA metrics to the Set, Set 2 and Set 3, which resulted from the described optimization routine. The objective of this analysis is to test the ability of the calculated ELA metric sets in distinguishing the 480 different function instances according to their categories (1- Separable Functions, 2- Functions with low or moderate conditioning, 3- Functions with high conditioning and unimodal, 4- Multi-modal functions with adequate global structure and 5- Multi-modal functions with weak global structure).

The comparison analysis of the sets was first conducted visually. Initially, the different sets was compared using density plots. This type of plot shows how each function category may be characterized according to the range of values of ELA metrics. Then, the different sets of ELA metrics is compared visually in a 2D scatter plot in the same way as similar works as in (MORGAN; GALLAGHER, 2017) and (MUÑOZ; SMITH-MILES, 2015). The choice for analysis in 2 dimensions is due to the simplicity of comparing function similarity (dissimilarity) by the distance of their representation dots in a scatter plot. Moreover, hidden structures, patterns, and clusters may be revealed by visual analysis of the 2D scatter plot. Also, because every run of the t-SNE technique produces a different embedded data, the resulting mean R_{NX} is also bound to vary. Therefore, the embedding procedure is repeated 25 times to account for the uncertainty of the quality metric.

In a second step, the Co-Ranking framework and one associated quality criteria metric is used to evaluate the quality of the embedded data. The goal of such analysis is to quantitatively compare the preservation of the data structure from the different set of ELA metrics. Also, a quantitative assessment is important because it shows how well one can validate what is represented in the lower dimensional space.

The package dimRed (KRAEMER *et al.*, 2018) is an R package that has built-in different dimensionality reduction algorithms and was adopted to perform the embeddings of the ELA metrics. The selected DR techniques are t-SNE and PCA due to their wide adoption and success in DR techniques. Also, it is interesting to compare their results since one is inherently restricted to a linear transformation of the data and the other is not.

The concept of the Co-Ranking Matrix is retaken in Figure 3.8. The general idea is to count the number of ranking order changes (swaps) and plot them in a matrix. The preservation of the ranking of data points in both higher and lower dimensional space indicates that the data structure has been preserved. If no ranking swaps have occurred, the Co-Ranking matrix should be formed by a single diagonal line.

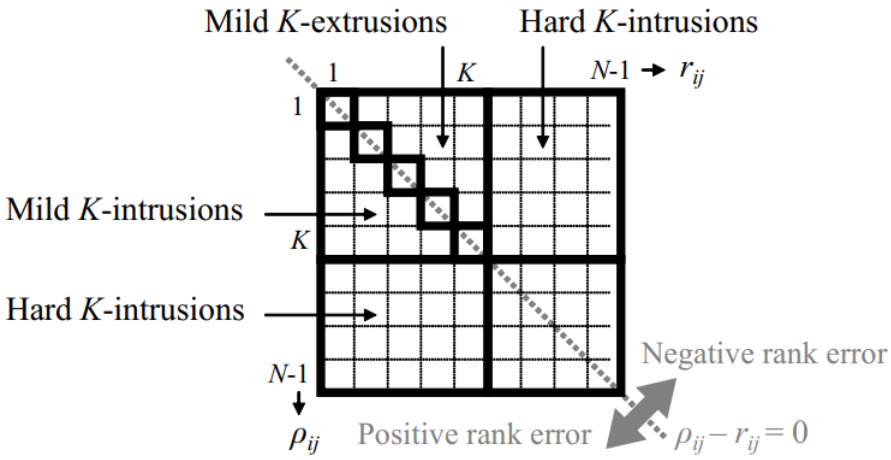


FIGURE 3.8 – Concept of a Co-Ranking matrix. From (KRAEMER *et al.*, 2018)

4 Results and Discussions

In this chapter, the results accomplished in the present dissertation are described. It follows the same sequence as established in Chapter 3: Materials and Methods. First, the selected ELA metrics is introduced. Also, the nature of each selected metric is discussed and compared to the sets of metrics from the ELA and ASP literature. In a second step, the visualization in a 2D scatter plot of the different sets of ELA metrics are displayed. The distance between the embedded data points serves as a measure of similarity between function instances. In a third step, a metric from the Co-Ranking framework is presented. This metric quantitatively describes the quality of the preserved data structure in a lower-dimensional space.

4.1 Selected Features

In this section, the selected features using the three proposed search strategies in Chapter 3: Materials and Methods are presented. They are introduced in Tables 4.1, 4.2 and 4.3 in association to which function characteristic they are able to measure. The nomenclature of function characteristics from (MERSMANN *et al.*, 2010) and (MALAN; ENGELBRECHT, 2013) shall be adopted to describe the selected features.

4.1.1 Set 1 - Unrestricted Size

Table 4.1 shows the resulting set of selected ELA metrics using the optimization procedure adopting an unrestricted size strategy. In this strategy, the optimization routine didn't incur any penalty in the number of the selected features. It is noticed that from a superset of 106 ELA metrics, a total of 49 metrics were selected. It can be noted that many of the selected ELA metrics are related to the same high and low-level properties of a function. For e.g, 6 metrics from the Local Search low-level property configure in the final set. The resulting mean R_{NX} after optimization is 0.649.

TABLE 4.1 – ELA Metrics in Set 1

High Level Property	**Low Level Property**	**Selected Metrics**
Global Structure Search Homogeneity Multimodality Basin Size Homogeneity	Convexity	Linear Deviation (lin_dev.orig) Absolute Linear Deviation (lin_dev.abs)
Plateaus Variable Scaling	Curvature	Minimum of Normalized Gradient (grad_norm.min) Mean of Normalized Gradient (grad_norm.mean) Upper Quartile of Norm. Gradient (grad_norm.uq) Number of NAs in Norm. Gradient (grad_norm.nas) Minimum of Scaled Gradient (grad_scale.min) Mean of Scaled Gradient (grad_scale.mean) Median of Scaled Gradient (grad_scale.med) Upper Quartile of Scaled Gradient (grad_scale.uq) Upper Quartile Hessian Matrix (hessian_cond.uq) Cond. Number Hessian Std.Dev (hessian_cond.sd)
Global Structure Search Homogeneity Multimodality	Level Set	MMCE_LDA_10 (mmce_lda_10) LDA_MDA_10 (lda_mda_10) QDA_MDA_10 (qda_mda_10) MMCE_QDA_25 (mmce_qda_25) MMCE_MDA_25 (mmce_mda_25) LDA_MDA_25 (lda_mda_25) QDA_MDA_25 (qda_mda_25) MMC_QDA_50 (mmc_qda_50)
Plateaus Global Structure Variable Scaling Separability Multimodality	Meta Model	Linear Model Ratio Max-Min Coeff. (lin_simple.coef.max_by_min) Linear Model Minimum Coeff. (lin_simple.coef.min) Quad. Model w/o Interaction R2 Adjusted (quad.simple.adj_r2) Quad. Model with Interaction R2 Adjusted (quad_w_interact_adj_r2)
Information and Disinformation		Max Entropy (ic.h.max) Max Epsilon (ic.eps.max) Partial Information Content
Global Structure Plateaus Multimodality	Y Distribution	Skewness
Global to Local Contrast Search Homogeneity Basin Size Homogeneity	Local Search	Best to Mean Contrast Original (best2mean_contr.orig) Best to Mean Contrast Ratio (best2mean_contr.ratio) Basin Sizes Average Non Best (basin_sizes_avg_non_best) Basin Sizes Average Worst (basin_sizes_avg_worst) Lower Quartile Function Evaluation (fun_evals.lq) Median Function Evaluation (fun_evals.median) Maximum Function Evaluation (fun_evals.max)

TABLE 4.1 – ELA Metrics in Set 1 - Cont.

Global Structure		Mean Dispersion Ratio 2% (disp.ratio_mean_02) Mean Dispersion Ratio 10% (disp.ratio_mean_10) Median Dispersion Ratio 5% (disp.ratio_median_05) Median Dispersion Ration 25% (disp.ratio_median_25) Mean Dispersion Difference 10% (disp.diff_mean_10) Mean Dispersion Difference 25% (disp.diff_mean_25) Median Dispersion Difference 25% (disp.diff_median_25)
Ruggedness		Correlation Length Fitness Distance Correlation Local Fitness Distance Correlation Global Skewness Length Scale Mode Length Scale Maximum Length Scale

4.1.2 Set 2 - Set with 10 Features

In the second proposed set of ELA metrics, its size is constrained to a cardinality of 10 features. As described in Chapter 3: Materials and Methods, this is the average number of features of the existing sets of ELA metrics from similar work that were adopted to be used in in the present work. Table 4.2 shows the resulting selected metrics. The resulting mean R_{NX} of this set is 0.651.

TABLE 4.2 – ELA Metrics in Set 2

High Level Property	**Low Level Property**	**Selected Metrics**
Plateaus Variable Scaling	Curvature	Median Cond. Number Hessian Matrix (hessian_cond.med)
Global Structure Search Homogeneity Multimodality	Level Set	MMCE_LDA_50 (mmce_lda_50)
Plateaus Global Structure Variable Scaling Separability Multimodality	Meta Model	Linear Model w/ Interaction R2 Adjusted (lin_w_interact.adj_r2) Quadratic Model Condition (meta.quad_simple.comd) Quadratic Model w/ Interaction R2 Adjusted (meta.quad_w_interact.adj_r2)

TABLE 4.2 – ELA Metrics in Set 2 - Cont.

Global to Local Contrast Search Homogeneity Basin Size Homogeneity	Local Search	Minimum Function Evaluations (local.fun_evals.min)
Information and Disinformation		Mean Dispersion Ratio 10% (disp.ratio_mean_10) Mean Dispersion Ratio 25% (disp.ratio_mean_25) Median Dispersion Ratio 2% (disp.ratio_median_02) Mean Dispersion Difference 2% (disp.diff_mean_02)

4.1.3 Set 3 - Set with 3 Features

The third proposed set was formed using a stratified initial population as described in Chapter 3: Materials and Methods. Table 4.3 lists the selected features. The resulting mean R_{NX} of this set following the described optimization procedure is 0.771.

TABLE 4.3 – ELA Metrics in Set 3

High Level Property	**Low Level Property**	**Selected Metrics**
Plateaus Variable Scaling	Curvature	Maximum Gradient Scaled (grad_scale.max)
Global Structure Search Homogeneity Multimodality	Level Set	MMCE MDA 50 (mmce_mda_50)
Information and Disinformation		Median Dispersion Ratio 2% (disp.ratio_median_02)

4.1.4 Qualitative Comparison of the Resulting ELA Sets

It is interesting to note that each of the resulting sets possesses metrics measuring different properties of the BBOB test functions. Besides, sets 1 and 2, generally present more than one single metric for a given function property.

This result is aligned with what the literature of ELA metrics has proposed, that is, the use of a "bag" of different metrics. In (MORGAN; GALLAGHER, 2017), the authors adopt metrics measuring Ruggedness and Neutrality (Fitness Distance Correlation and Correlation Length), Information and Disinformation (Dispersion and Information Content). Also, in (MORGAN; GALLAGHER, 2017) they propose a set of metrics based on statistics (Mean, Median, Mode, Maximum, Minimum, Kernel Bandwith, and Entropy) of a new ELA metric, which they call Length Scale, which measures a function Neutrality and Ruggedness. In (SUN *et al.*, 2014), the authors propose a set of metrics related to Function (Y) Statistics (Number of Peaks, Skewness, Kurtosis and Number of Dimensions), Ruggedness and Neutrality (Median Length Scale) and Information and Disinformation (Information Content, Partial Information Content, and Information Stability). Finally, in (MERSMANN *et al.*, 2010), the authors use a set including metrics related only to the Level Set of a function, which is, in turn, related to the Global Structure, Search Homogeneity and Multimodality of a function.

Metrics related to a function gradient and its dispersion seem to be important since, in each of the proposed sets 1, 2 and 3, there is a representative ELA metric related to these characteristics. As established in Chapter 2: Literature Review, the metrics measuring the dispersion of a function calculate the average pairwise distance of the top-ranked points of a sample, where this rank is formed by the function value below a given threshold (2%, 5%, for example). Also, because the gradient measures the direction of the fastest increase of a function, it is not a surprise that related metrics were included in the proposed different sets.

4.2 Visual Analysis of the Test Set of Functions

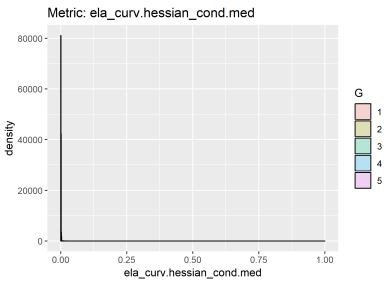
In the present section, the different sets of ELA metrics are used to distinguish the function categories. In the first step, the density plots of each metric included in the set are displayed. This exploratory analysis is interesting because it shows the probability of a function belonging to a given category given an ELA metric value. In a second step, a 2D scatter plot shall be displayed. The scatter plots depict the similarity or dissimilarity between data points according to the distance of the dots in the plot.

4.2.1 Analysis of Density Plots

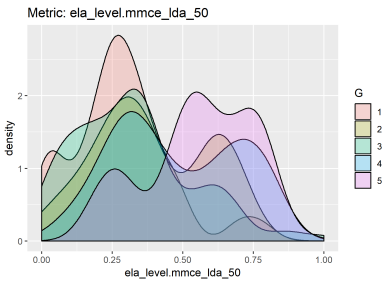
In order to present a more parsimonious subsection (not polluted with plots), we hereby present only the density plots of Set 2, which includes 10 ELA metrics. The density plots of all subsets are shown on Appendix E. All density plots from the 7 subsets lead to similar conclusions.

4.2.1.1 Proposed Set 2 - ELA 10 Metric

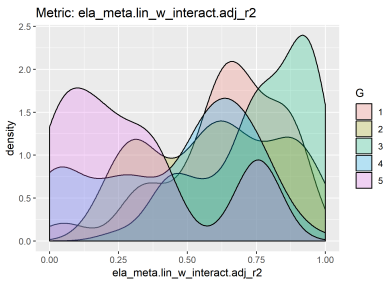
Figures E.6 and E.7 depict the normalized values of the ELA metrics in Set 2, along with their associated density values. It is noticeable from plots (d)-(g) on Figure E.7 that the metrics associated with the dispersion of a function instance cannot to used individually to classify function categories. This is evidenced by the overlay of colors. Also, some ELA metrics present similar value to all function categories, such as the Quadratic Model Condition in Figure E.6, where a single vertical line is depicted on the plot.



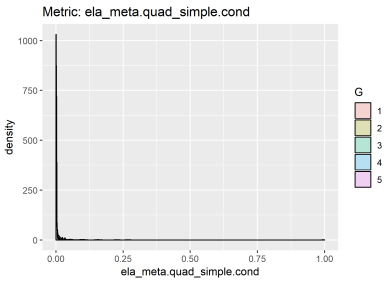
(a) Median Hessian Matrix Cond.



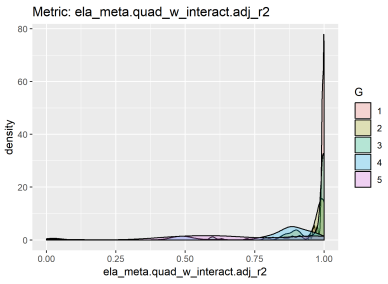
(b) MMCE-LDA-50



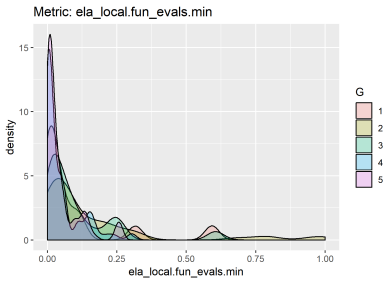
(c) Linear with Interaction Adj-R2



(d) Quadratic Model Condition

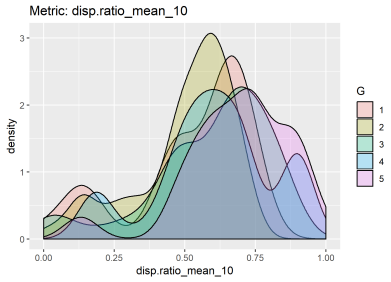


(e) Quadratic with Interaction Adj-R2

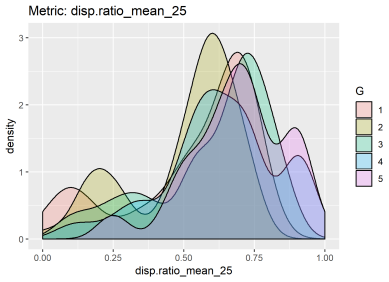


(f) Min Function Evaluations

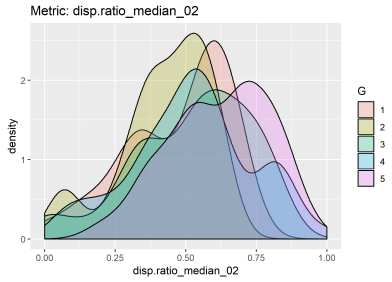
FIGURE 4.1 – ELA Metrics in Set 2 - A



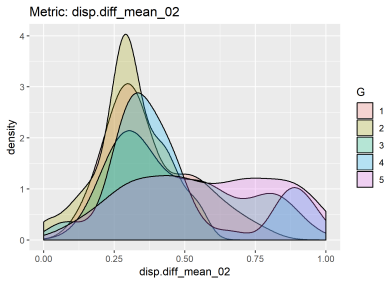
(a) Mean Dipersion Ratio 10%



(b) Mean Dispersion Ratio 25%



(c) Median Dispersion Ratio 2%



(d) Mean Dispersion Diff. 2%

FIGURE 4.2 – ELA Metrics in Set 2 - B

4.2.1.2 Conclusion on the Density Plots

By the examination of the different density plots in subsubsection 4.3.1.1 and Appendix E, one may reach some conclusions. First, it becomes evident the need for a set of metrics approach when characterizing the selected continuous functions in the BBOB09 benchmark. For a given specific value of an ELA metric, seldom a function category could be easily classified. Rather, more often, for a given specific value of an ELA metric, 2 or more categories presented similar density value.

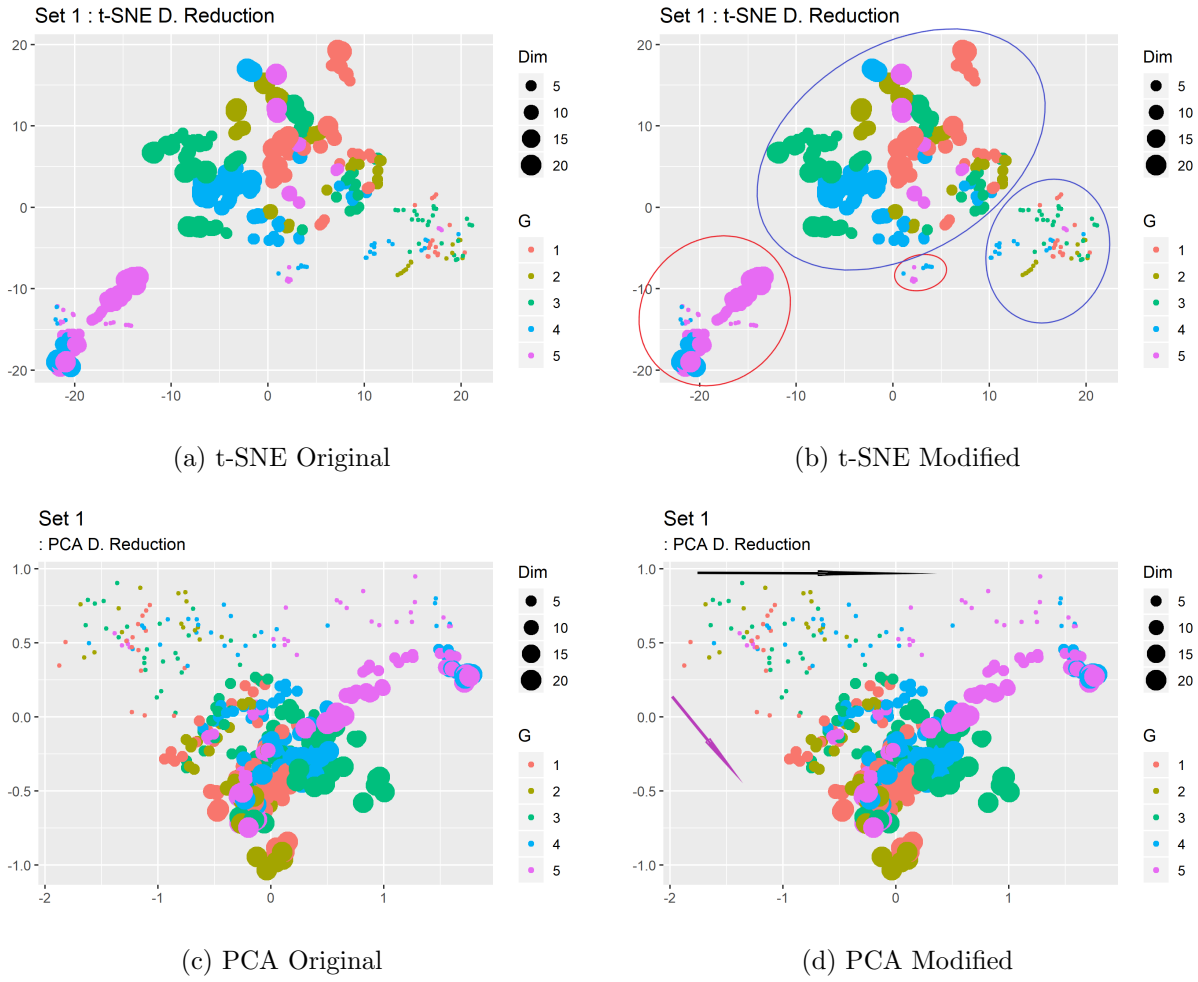
Also, it is evident that some function instances from the same category presented multiple modes of values. This result may imply the existence of subgroups of similar functions instances within the same function type. On the other hand, some sets possess ELA metrics where all functions presented very similar value. In such cases, this metric doesn't add information for distinguishing between categories and thus may be even removed from the set.

4.2.2 Analysis of 2D Scatter Plots

In this subsection, the 2D scatter plots of embedded data are displayed. As described in Chapter 3: Materials and Methods, each set of ELA metrics is represented in a lower (2D) dimension to depict the similarity between function instances.

The embedded 2D plots for the PCA and t-SNE techniques are displayed in the original form and a modified one, where arrows and circles may facilitate their interpretation. Each function category is plotted in a different color. Also, the dimension of a datapoint is represented by the size of a point in the plot. The coordinates x-y in the plots represent each coordinate in the 2D (x1, x2) embedded space.

4.2.2.1 Set 1 Plots



(a) t-SNE Original

(b) t-SNE Modified

(c) PCA Original

(d) PCA Modified

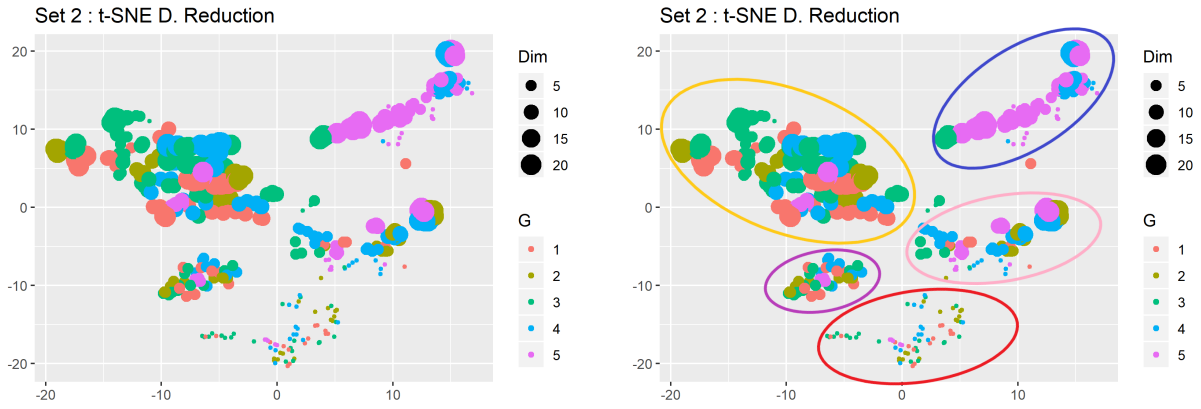
FIGURE 4.3 – 2D Scatter Plots for Set 1

Some patterns may be recognized by inspection of the embedded data of set 1. One shown in Figure 4.3 (b) is that functions of categories 4 and 5 (Multi-Modal with adequate

global structure and Multi-Modal with weak global structure, respectively) are clustered together, depicted in the red ellipses. Also, function instances in categories 1, 2 and 3 (Separable, Low or Moderate Conditioning and High Conditioning, respectively) are grouped, shown by ellipses in blue. It seems that functions with low dimensionality are pictured in the lower right corner of the 2D space.

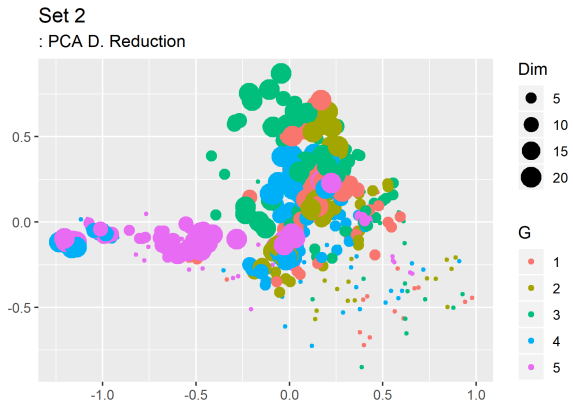
The PCA embedded plot in Figure 4.3 (c) and (d) also depicts patterns. The purple arrow seems to indicate the direction of higher-dimensional data points, whereas the black arrow seems to indicate the direction for separation of function categories (E.g: Functions in categories 4 and 5 occupy the right corner of the plot in comparison of functions in category 3, which are found in the center of plot).

4.2.2.2 Set 2 Plots



(a) t-SNE Original

(b) t-SNE Modified



(c) PCA Original

(d) PCA Modified

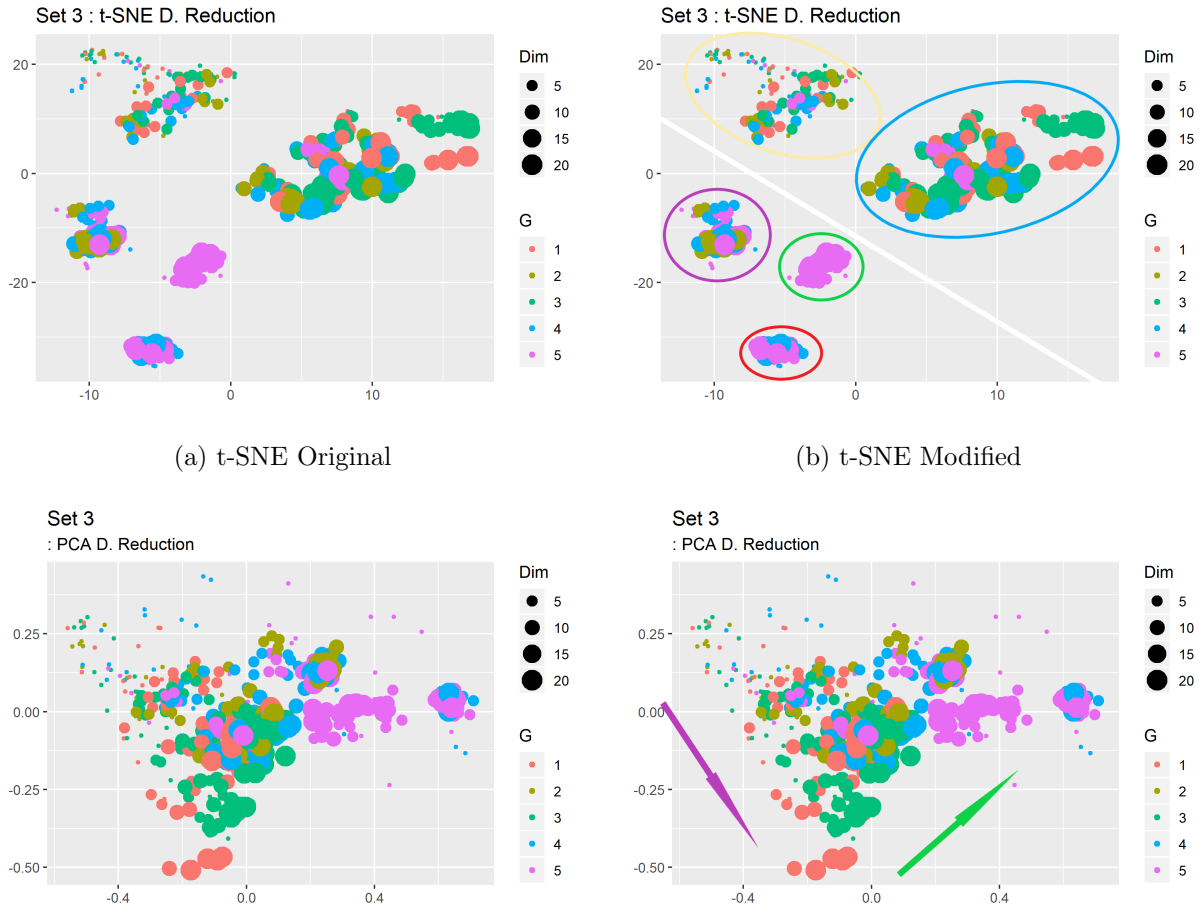
FIGURE 4.4 – 2D Scatter Plots for Set 2

Figure 4.4(b) depicts the embedded via the t-SNE technique. Different clusters can be seen inside the colored ellipses. Also, functions in categories 4 and 5 tend to be clustered

on the right side of the plot. The dimensionality of functions seems also to increase from the bottom part of the plot to the upper part.

In plot (d) PCA DR technique also depicts some patterns. The black arrow indicates the direction of functions in categories 4 and 5. The red arrow indicates a direction of growth of the dimensionality of the functions.

4.2.2.3 Set 3 Plots



(a) t-SNE Original

(b) t-SNE Modified

(c) PCA Original

(d) PCA Modified

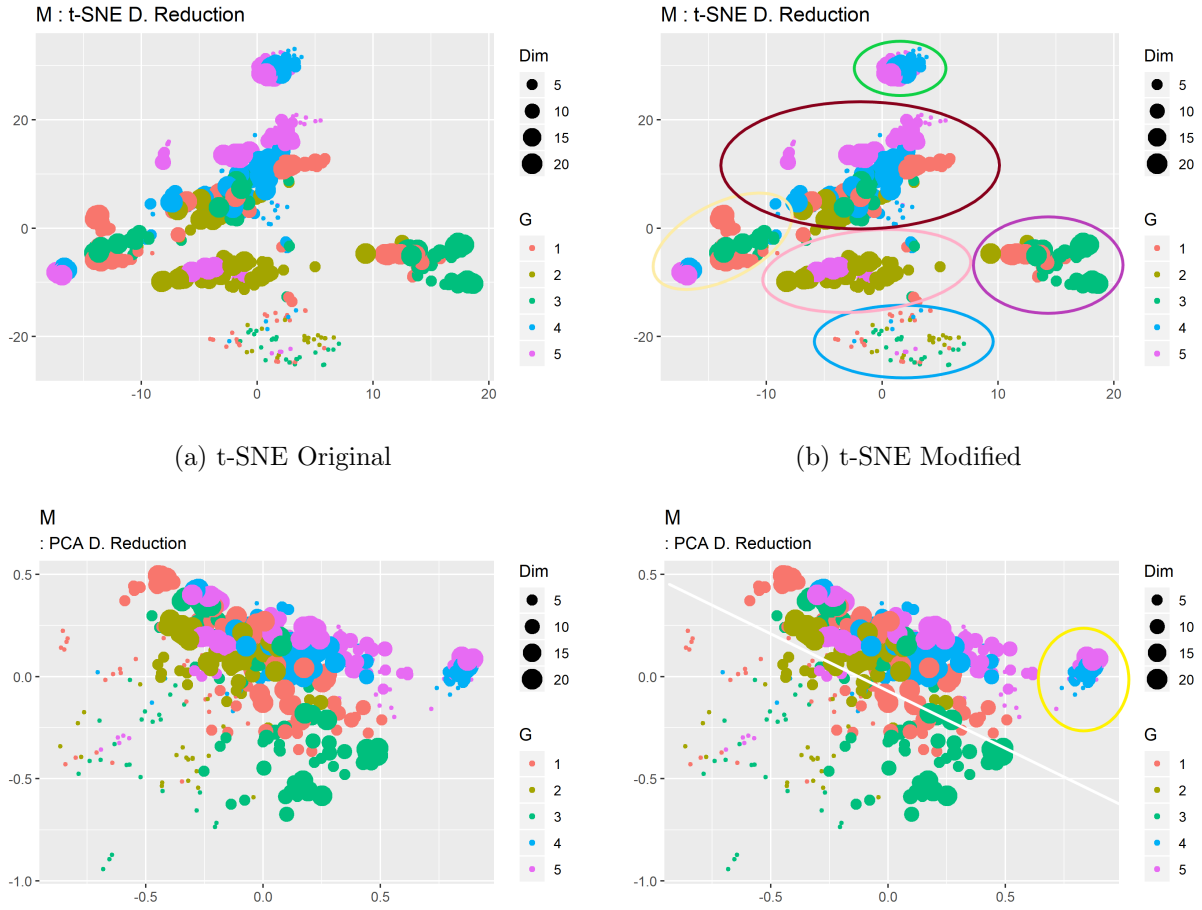
FIGURE 4.5 – 2D Scatter Plots for Set 3

In Figure 4.5 (b), the embedded data with the t-SNE technique is presented. Functions in categories 4 and 5 tend to be clustered in the lower-left corner of the plot, in groups inside colored ellipses. On the other hand, functions of group 1, 2, 3 and some representatives in category 4 are clustered in the upper, right corner of the plot. Lower dimensional functions are also grouped in the top left corner of the plot in general.

The PCA technique also manages to capture patterns. The purple arrow in Figure 4.5 (d) shows a probable direction of increase in the dimensionality of data, whereas the

green arrow depicts a direction in the change of group type.

4.2.2.4 Set M Plots



(a) t-SNE Original

(b) t-SNE Modified

(c) PCA Original

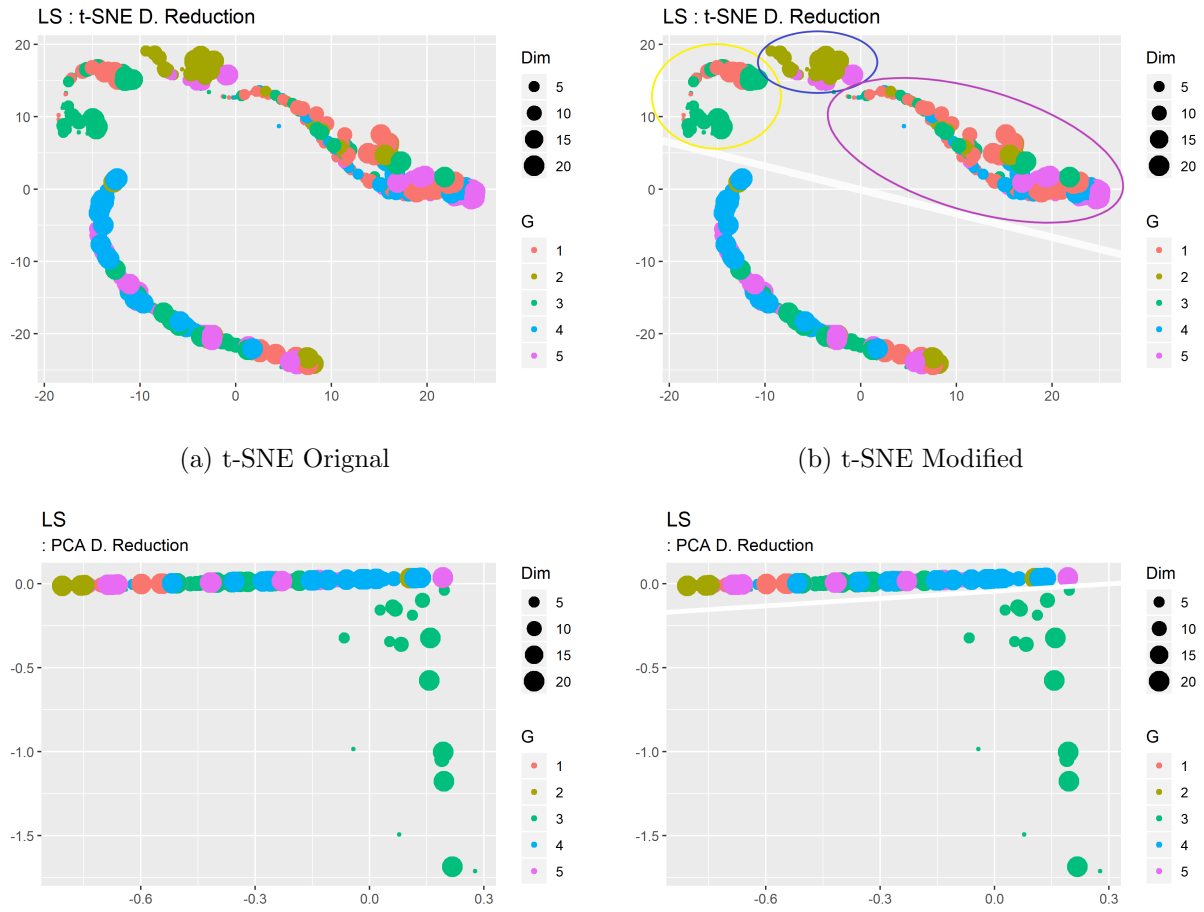
(d) PCA Modified

FIGURE 4.6 – 2D Scatter Plots for Set M

Figure 4.6 displays the embedded data using set M. In plot (b), the t-SNE DR technique grouped different functions from different categories inside the depicted colored ellipses. Also, data points (functions) from lower dimensions are depicted lower, the bottom part of the plot.

Differently, the PCA technique doesn't show distinctively clusters of data points by the exception of some instances of functions in categories 4 and 5, inside the yellow ellipse. The data points are rather depicted in a "cloud" of points. However, a rather distinct pattern can be inferred from the plot (d). Below the white diagonal line, there is a higher concentration of functions with lower dimensions mainly from categories 1,2 and 3, whereas above the line there is a concentration of functions from categories 4 and 5.

4.2.2.5 Set LS Plots



(a) t-SNE Orignal

(b) t-SNE Modified

(c) PCA Original

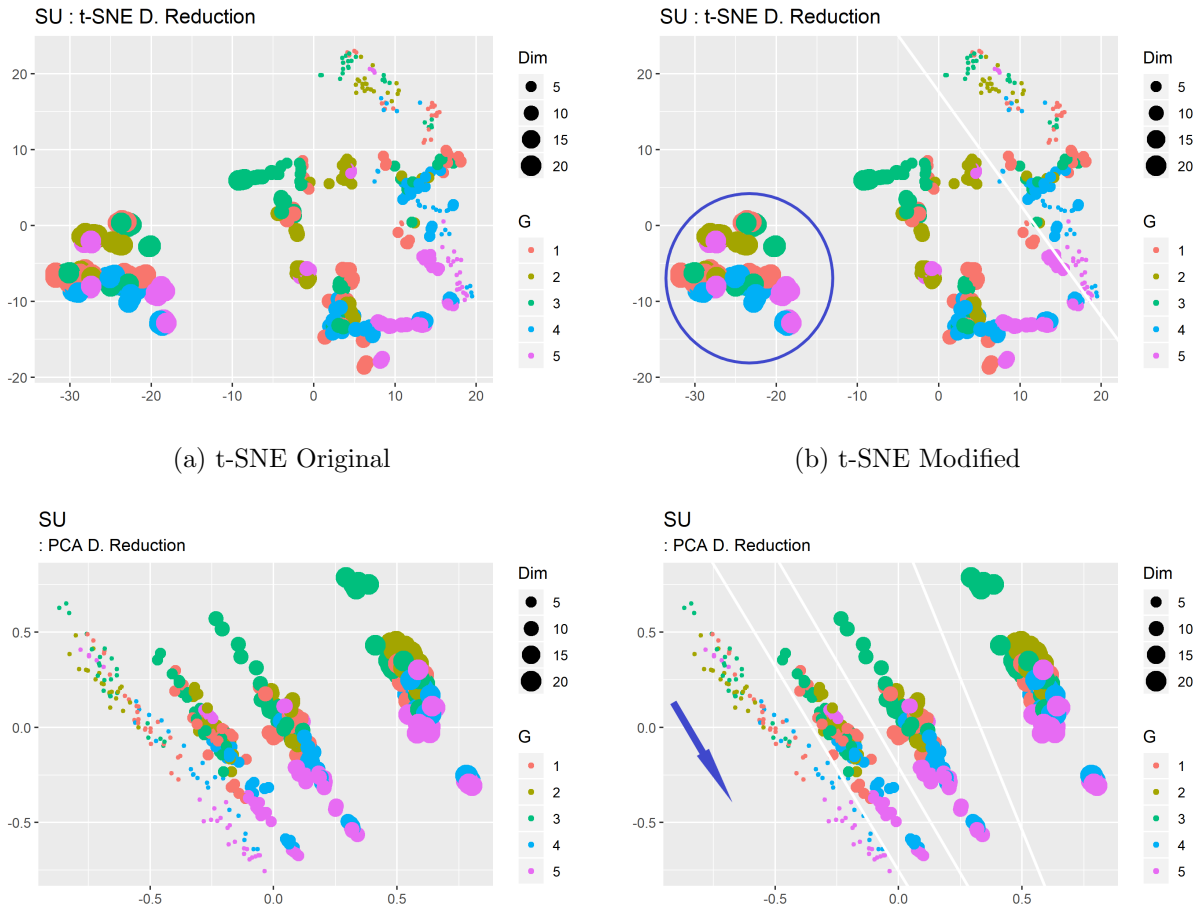
(d) PCA Modified

FIGURE 4.7 – 2D Scatter Plots for Set LS

Figure 4.7 depicts the embedded data using set LS. In plot (a), the t-SNE technique doesn't seem to depict clear patterns on the data. A white line was drawn in Figure 4.7 (b) to separate two distinct groups of data. Generally, the functions in category 4 are found below the drawn line, whereas functions in category 1 are depicted above the line.

Similarly, the embedded data using the PCA technique, as depicted in Figure 4.7 (d) doesn't provide too much information on clusters of data. Below the white line, there is a higher concentration of functions of group 3, which are separated from the rest of the functions from different groups, above the drawn line.

4.2.2.6 Set SU Plots



(a) t-SNE Original

(b) t-SNE Modified

(c) PCA Original

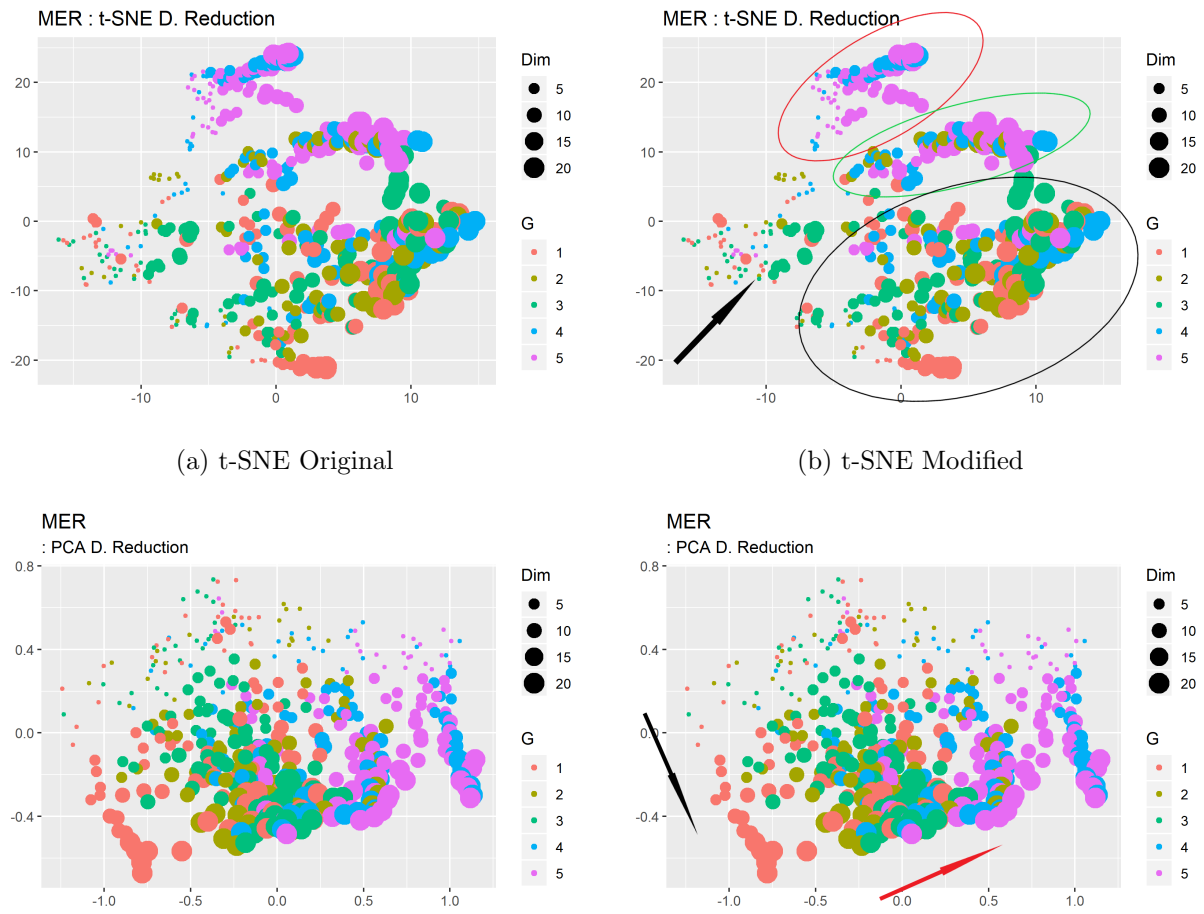
(d) PCA Modified

FIGURE 4.8 – 2D Scatter Plots for Set SU

Figure 4.8 depicts the embedded data using set SU. In plot (a) it is noticeable a higher concentration of lower-dimensional points above the white line. Moreover, the dimensionality of the data seems to increase in the direction of the lower corner one the plot.

The embedded data by the PCA technique shows an interesting pattern. First, the embedded data is separated by their dimension, as shown in white lines cutting the plot. Moreover, the same pattern of colors seems to be preserved across each group (functions of category 3 in one end and functions from categories 4 and 5 in the other end of the cloud of points). This pattern is shown as indicated by the blue arrow.

4.2.2.7 Set MER Plots



(a) t-SNE Original

(b) t-SNE Modified

(c) PCA Original

(d) PCA Modified

FIGURE 4.9 – 2D Scatter Plots for Set MER

Figure 4.9 depicts the embedded data using set MER. In plot (b) it is noticeable different categories of function instances inside colored ellipses. Besides, functions in categories 4 and 5 tend to be depicted in the upper half of the plot, whereas functions in 1 and 2 in the bottom half of the plot. The black arrow also indicates a direction in the increase of dimensionality of the data.

The embedded data using the PCA DR technique also shows patterns. The red arrow indicates the direction of function category. There is a higher concentration of functions of categories 4 and 5 in the right corner of the plot. Instances of group 1 are concentrated on the left side and in the middle of the plot, whereas function instances belonging to category 3 are located in the middle of the plot. The black arrow indicates the direction of the increase of the dimensionality of the data.

4.2.2.8 Scatter Plot Comparison

Overall the different set of ELA metrics displayed function similarity and dissimilarity in a lower-dimensional space. With some sets, their associated 2D plots produced more visually identifiable patterns. For instance, set 1 in Figure 4.3, set 2 in Figure 4.4, set 3 in Figure 4.5 and set 4.8 depict clusters - evidenced inside colored ellipses - of similar functions according to the metrics included in each set. However, metrics in set LS embedded by both PCA and t-SNE techniques depicted in Figure 4.7 don't seem to produce much information on the data points.

In general, the 2D scatter plots reveal interesting information that seems to repeat in the different embedded 2D plots. Across plots from different ELA sets, for example, some function instances from category 4 and 5 are often clustered. These are the functions that present characteristics such as Multi-Modularity with global structure and Multi-Modularity with weak structure. However, not all instances of functions from category 4 were depicted close to functions from category 5. As shown for example in Figure 4.5 (d) and in Figure 4.6 (b), some instances in category 4 were depicted rather distant from function instances from category 5. Moreover, many function instances from categories of rather different characteristics are often depicted close together. For example, instances of functions from category 1 and 3 (Separable Functions and Functions with High Conditioning and Unimodal, respectively), are often depicted closer (colors red and green), in spite of belonging to categories with distinct characteristics. This result may be evidence that the BBBOB09 function categories as defined in (HANSEN *et al.*, 2009) serve rather as a theoretical guide for classification of functions. In practice, functions belonging to different categories may possess similar characteristics and be depicted close to each other as seen in the 2D embedded data. One possible reason for such finding is that despite belonging to different categories, those specific function instances share similar underlying topologies (characteristics) in such point of the space. In (MORGAN; GALLAGHER, 2017) the authors reach similar conclusions.

Finally, despite the use of DR techniques such as t-SNE and PCA provides useful information about the similarity between function instances, the analysis of the 2D plot must be done with caution. DR techniques employ transformations in the data so it can be displayed in a lower-dimensional plot. It is reasonable to expect that some deformations and noise be introduced in the final result. The 2D plots don't show, nor quantify which portions of the space have closer resemblance with the original, higher-dimensional data. For this, specific metrics that quantify the information preservation of the original data must be inspected. Next section displays the result analysis of the data using the Co-Ranking framework.

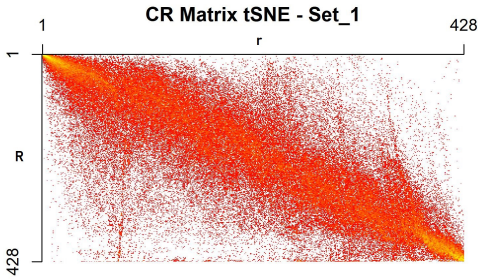
4.3 Quantifying Information Preservation with the Co-Ranking Framework

In this section, the Co-Ranking framework is used to assess the preserved information of the ELA sets when the DR techniques were employed. The use of the Co-Ranking framework in the present work shall provide in a quantitative aspect the reliability of the information provided from the embedded data. Each entry in the matrix, is depicted by a red dot and corresponds to the count number of rank swaps from a rank R (in the high-dimensional space) to rank r (in the lower-dimensional space).

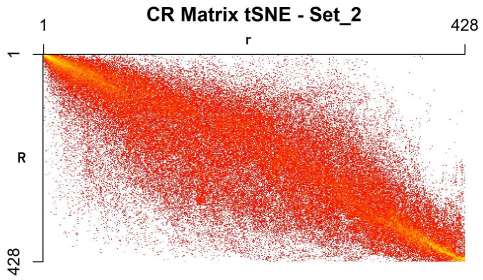
4.3.1 The Co-Ranking Matrix

4.3.1.1 Assessment of t-SNE Embedding

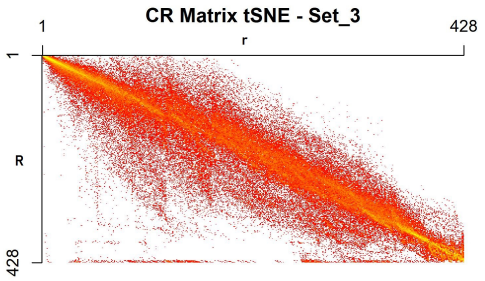
Figure 4.10 and Figure 4.11 display the resulting Co-Ranking matrices by embedding the different sets of ELA metrics with the t-SNE technique.



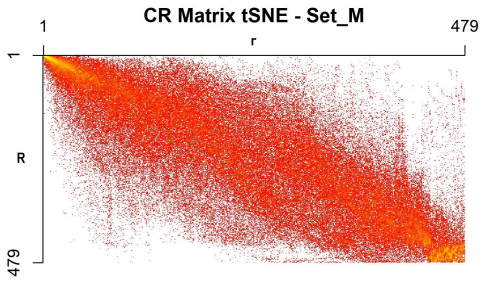
(a) CR Set 1



(b) CR Set 2

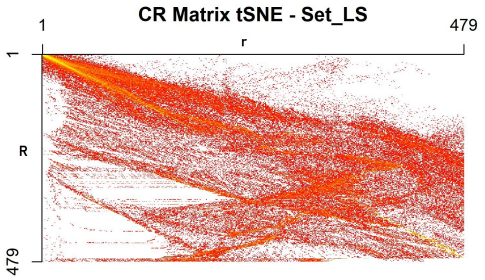


(c) CR Set 3

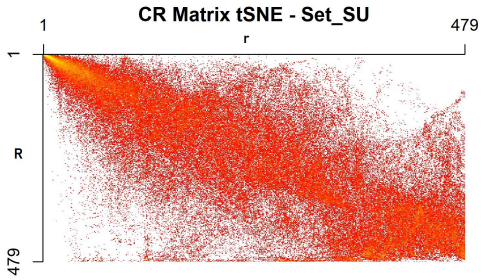


(d) CR Set M

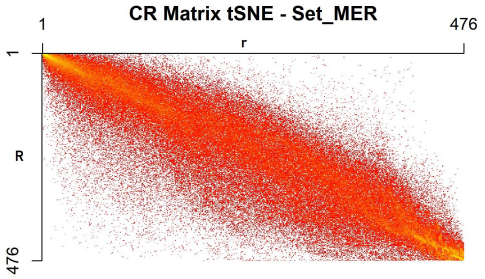
FIGURE 4.10 – Resulting Co-Ranking Matrices with t-SNE - A



(a) CR Set LS



(b) CR Set SU



(c) CR Set MR

FIGURE 4.11 – Resulting Co-Ranking Matrices with t-SNE - B

By inspection of the resulting Co-Ranking matrices, one may conclude the embedded data is far from optimal. As established throughout the present work, the Co-Ranking matrix displays the number of rank swaps, that is, it counts the number of neighborhoods in the original higher-dimensional data of a given size i that, when suffers a transformation by an embedding procedure, changes (swaps) the size of rank i to a new rank of size j.

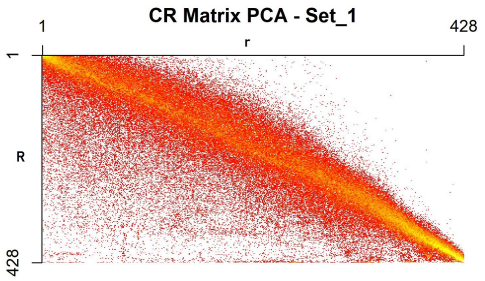
It is clear from the plots that the t-SNE DR technique introduces nonnegligible distortions in the resulting projected data in the 2D space. Many non zero entries above the diagonal line indicate that the projection depicted data points more distant than they are in the original data set (increase in their rank). Similarly, the t-SNE DR also depicts points closer than their original structure (decrease in rank).

A visual analysis shows that Set 3, Set MER and Set 1 are sets that present fewer distortions in their associated embedded data. Sets LS and SU present a high number of intrusion (data points become closer in lower dimension) and of extrusion (data points become distant in lower dimension) which is evidenced by the number of entries below and above the diagonal line in their associated Co-Ranking matrix.

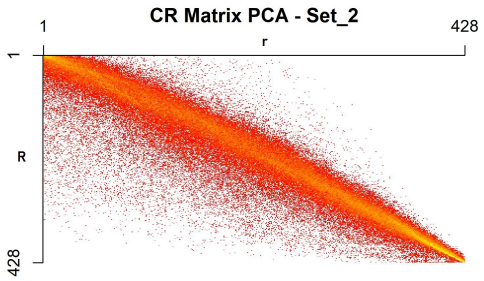
4.3.1.2 Assessment of PCA Embedding

Figure 4.12 and Figure 4.13 show the Co-Ranking matrices resulting from PCA DR technique. Overall, the resulting matrices indicate much better performance. It can be visually noted that there was much less extrusion and intrusions of data points since there are more entries in the matrices confined around their diagonal area.

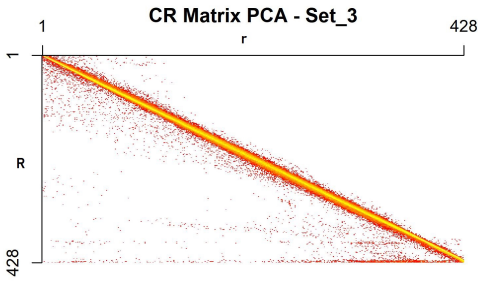
Also, the PCA technique seems to cause more distortion of the type intrusion than extrusion in the embedded data. An inspection of the 7 Co-Ranking plots shows a higher concentration of entries below the diagonal area. This indicates that the PCA techniques tend to depict data points closer than they are, thus reducing the size formed by their neighborhoods (ranks).



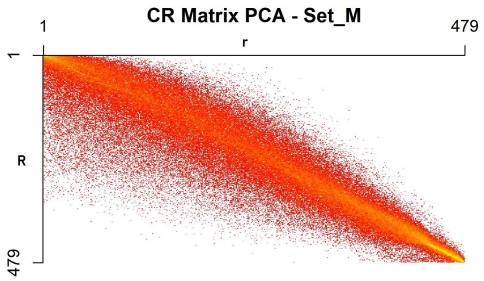
(a) CR Set 1



(b) CR Set 2

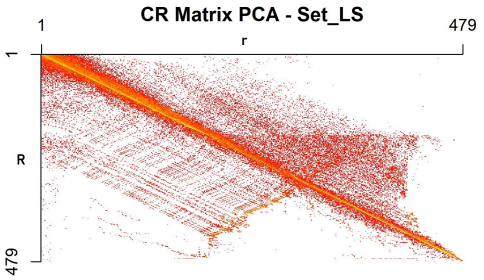


(c) CR Set 3

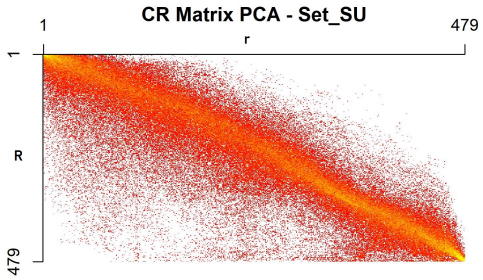


(d) CR Set M

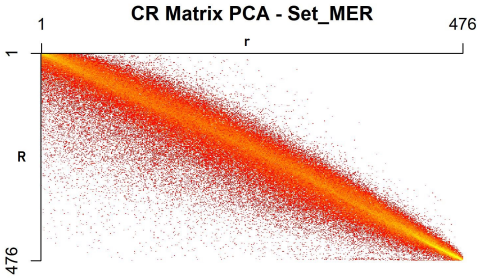
FIGURE 4.12 – Resulting Co-Ranking Matrices with PCA - A



(a) CR Set LS



(b) CR Set SU



(c) CR Set MER

FIGURE 4.13 – Resulting Co-Ranking Matrices with PCA - B

Finally, Set 3, Set LS and Set MER perform very well. The resulting embedding of Set 3 is near-optimal since there are very few entries off the diagonal line in its associated Co-Ranking matrix. Also, whereas the embedding of Set LS with the t-SNE technique produced a lot of distortion, PCA seems to preserve rather well the distance between points. In this sense, one may conclude that different DR techniques may perform better or worse in representing an original, higher-dimensional data set in a lower dimension.

4.3.2 Analysis of the R_{NX} Metric

4.3.2.1 Global Assessment - Mean R_{NX}

A visual inspection of the Co-Ranking matrix produces insight into the overall quality of the embedding. However an analysis based solely on the heatmaps remains qualitative, as it doesn't quantify how much more effective one given embedded data is compared to another. To achieve a quantitative comparison of the preserved information (quality) of the embedded ELA sets, the mean R_{NX} metric was adopted. Table 4.4 resumes its mean value calculated from a sample of 25 repetitions since the t-SNE technique possesses a

probabilistic nature and each embedding produces a different result.

Set ID	**Mean (Mean R_{NX} - tSNE)**	**Mean (Mean R_{NX} - PCA)**
Set 1	0.627	0.675
Set 2	0.655	0.804
Set 3	0.758	0.943
Set M	0.606	0.763
Set LS	0.531	0.860
Set SU	0.531	0.689
Set MER	0.691	0.806

TABLE 4.4 – Mean R_{NX} from a sample of size 25

An inspection of the second column of Table 4.4 confirms the embeddings with the t-SNE DR technique is far from optimal. For instance, the highest value of this criterion is 0.758, resulting from an embedding with Set 3. The interpretation of this result is as follows: A value of 0 indicates a random embedding inside a k-ary neighborhood, whereas a value of 1 indicates a perfect embedding. Therefore, a value of 0.758 means approximately 76% of all neighborhoods (of size 1, 2, to...N-1, where is N the number of data points), correctly preserve their size when embedded in the lower, 2D space. Another way of interpreting this result is with an error rate manner. That is, approximately 24% of all the neighborhoods suffered distortion when the data was embedded in a 2D space. The worst embedded data are the ones from Set SU and Set LS, where approximately 47% of the projected manifold suffered distortion.

Differently, embedded data using the PCA technique shows better performance. For instance, an embedding using Set 3 provides a mean value of 0.943 of the mean R_{NX} metric, that is, about 6% of the embedded manifold present distortion. Set 2, Set LS and Set MER also present high mean R_{NX} values (above 0.8). It is also interesting to note the shift in the quality of the embedded data using set LS. Using the PCA technique over t-SNE generates an improvement of about 62% in the quality of the embedding. In this sense, it becomes clear that according to the structure of the dataset (manifold), some DR techniques may be more or less suitable. Figures 4.14 and 4.15 show the comparison of the mean R_{NX} resulted from the application of the t-SNE and PCA techniques.

Although the Mean R_{NX} metric encapsulates in a single value the overall quality of the embedding, its interpretation must be used with caution. A mean R_{NX} value of 0.60, such as the one provided by Set M in conjunction with the t-SNE indicates that 40% of the embedded manifold suffers distortion. The metric, however, doesn't quantify the magnitude of the distortion. That is, a rank swap from 1 to 49 (very strong extrusion)

counts the same as a rank swap from 1 to 10 (mild extrusion) in a dataset comprised of 50 data points, for example. Also, the Mean R_{NX} calculates the mean over the entire set of R_{NX}, thus obscuring which size of neighborhoods are better represented in the lower dimensional space.

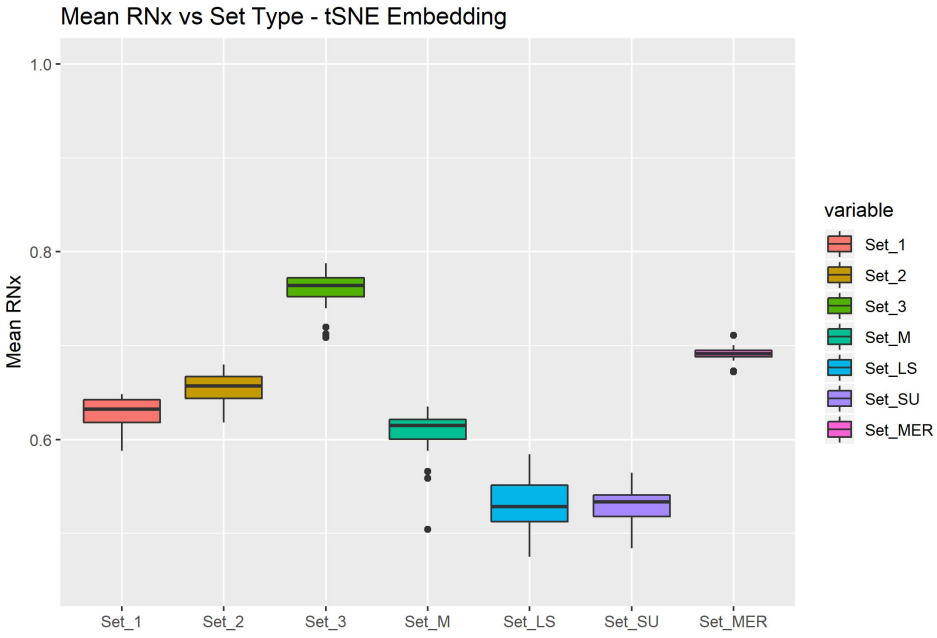


FIGURE 4.14 – Resulting Mean R_{NX} using t- SNE

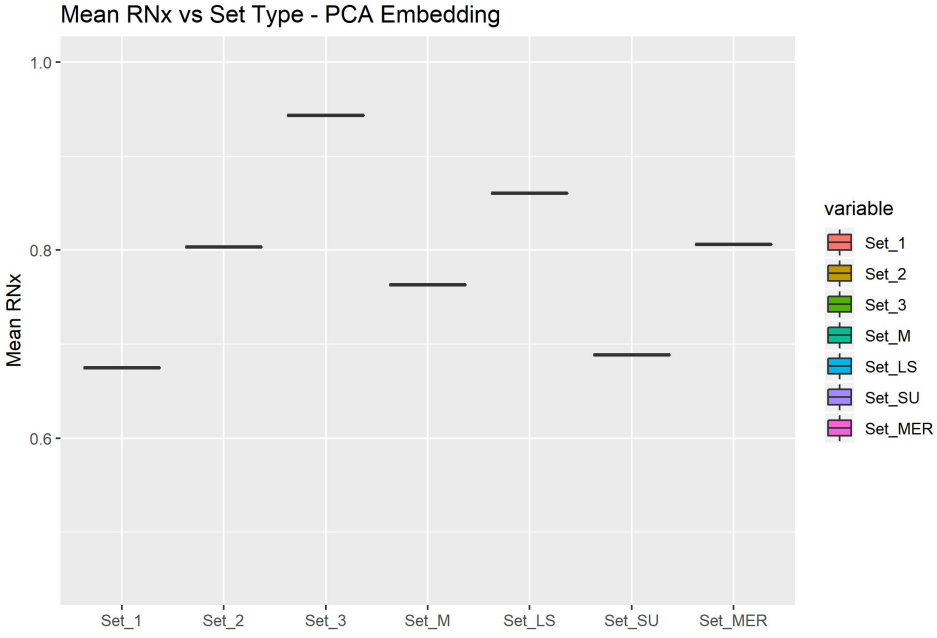
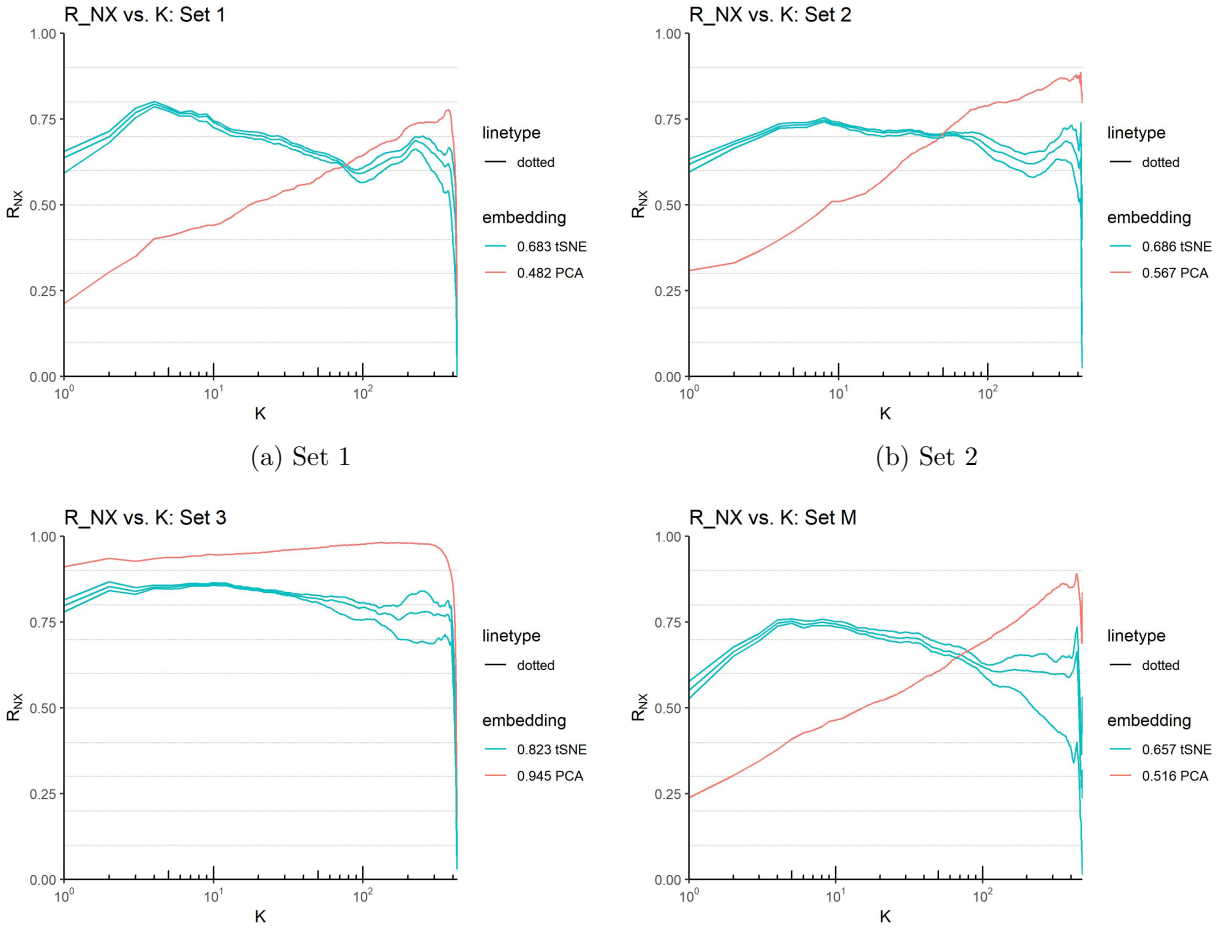


FIGURE 4.15 – Resulting Mean R_{NX} using PCA

4.3.2.2 Local Assessment - R_{NX} vs K

The mean R_{NX} metric encapsulates the overall quality of the embedded data into a single figure. However, it might be also interesting to asses how the quality of the manifold varies according to the size of the neighborhood k. This information may be useful in combination with the embedded data in 2D plots, where the metric may serve as a trustworthiness estimate of what is being depicted. For example, an R_{NX} value of 0.50 for a neighborhood of size $k = 4$ has the following interpretation: 50% of the neighborhoods of size 4 suffered some distortion. One may even go further and use this information in the 2D plot, where given a depicted point in the embedded space and its associated neighborhood of size 4, there is a 50% probability that points depicted inside of the given neighborhood doesn't reflect its true neighborhood of same size in the original, higher-dimensional space.

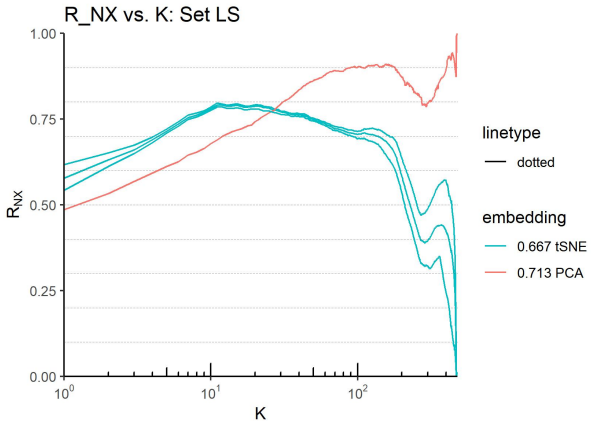
Figure 4.16 and 4.17 depict the R_{NX} vs K plots for the different ELA sets.



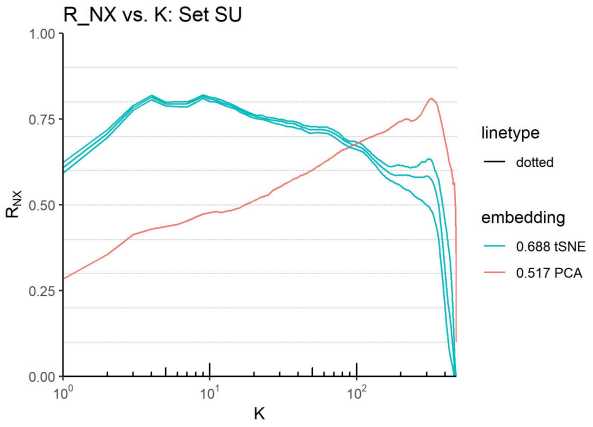
(a) Set 1 (b) Set 2

(c) Set 3 (d) Set M

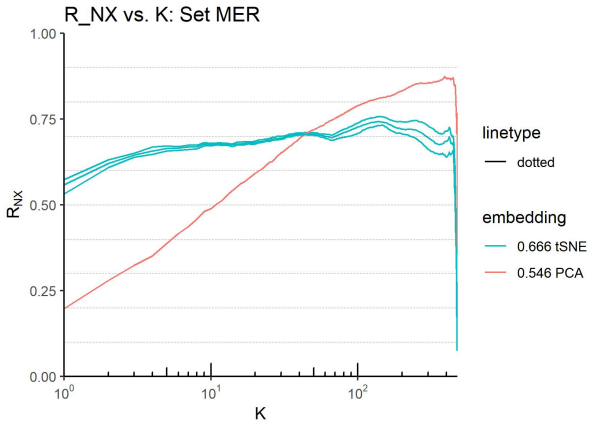
FIGURE 4.16 – R_{NX} vs K Plot - A



(a) Set LS



(b) Set SU



(c) Set MER

FIGURE 4.17 – R_{NX} vs K Plot - B

An inspection of Figures 4.16 and 4.17 produces interesting insights. First, it is noticeable the change in the R_{NX} values according to the neighborhoods of sizes K and the employed DR techniques, that is, t-SNE and PCA. This finding sheds some light on the suitability and limitations of a particular technique for embedding a given set. For instance, in Figure 4.16 (a), the embedding produced by t-SNE technique produces a maximum R_{NX} peak of 0.77 with a neighborhood of size k = 3. Beyond this neighborhood size, the embedded manifold continuously suffers distortion, shown by the decreasing values of the R_{NX} metric. Differently, as depicted in Figure 4.17 (b), the embedding of Set 1 with the PCA technique produces a manifold of increasing resemblance to the true manifold with the increase of the neighborhood size k. The result indicates that t-SNE performs better in preserving local neighborhoods, while PCA technique provides better results when the size of the neighborhood increases.

Figure 4.16 (c) shows the embedding using Set 3 outperforms all other ELA metric sets when the criterion is the preservation of the manifold. The PCA technique produces R_{NX} values close to 0.95 in small and large neighborhoods. The same behavior is repeated

with the t-SNE technique, though it presents a lower mean R_{NX} value compared to PCA.

4.4 Summary

This chapter has presented the results achieved in this work. The main points can be summarized as follows:

I. The proposed ELA metrics in Set 1, Set 2 and Set 3 cover different continuous function characteristics analogous to proposed ELA metric sets in the literature.

II. It is difficult to use a single ELA metric to discriminate between function groups since many of these groups possess representative instances with very similar values.

III. Clusters of similar functions were depicted in the 2D plot. However, data points inside a cluster don't necessarily belong to the same function type. This is in agreement with previous findings in the ELA metric research.

IV. Embedding sets of ELA metrics produce distortions in the original data structure (manifold). Therefore, function instances may be depicted closer (or farther) than in the original, higher dimensional dataset. This, in turn, may induce false conclusions about the similarity between functions.

V. Some combinations of sets of ELA metrics and an adequate DR technique may produce less distortion in the resulting manifold.

VI. Some DR techniques are more suitable for representing small neighborhoods (t-SNE in the present work), whereas others perform better in representing the global structure of the manifold.

5 Conclusion

The field of Optimization has seen a huge increase in the number of algorithms. Not only researchers developed new metaheuristics based in some nature metaphor, but they could also combine existing techniques to form new ones in a hybrid scheme. As a result, besides having an optimization problem to solve, researchers face the challenge of carefully selecting an adequate method from a myriad of choices. Therefore the Algorithm Selection Problem has become a more active field of research, whereas the feature space a central part of the problem as seen in (RICE, 1976).

The present dissertation belongs to this category of research. By comparing different sets of ELA metrics, and, employing Dimensionality Reduction techniques, the main objective was to select the metrics which best preserved the information of the higher-dimensional manifold in a 2D space. The embedding of metrics related to the characteristics of continuous functions is important because it renders visual interesting information that may not be captured from a complete table of measured figures.

The results presented are positive to the field. They will be further detailed in the next subsection.

5.1 Main Contributions

The author believes that the present work has successfully achieved its objective. They may be summarized as follows:

- Proposal of New Sets of ELA Metrics: With the adoption of three different strategies for an optimization routine, three accompanying sets of ELA metrics were produced (Set 1, Set 2 and Set 3). Though by the inspection of the mean R_{NX} metric, Set 1 and Set 2 aren't much more performing than the established sets in the literature, that is not the case with Set 3. Set 3 is the best performing set concerning the mean R_{NX} metric. Also, because the set cardinality is only 3, one might argue that reducing a manifold from 3D to 2D may cause fewer distortions than performing an embedding from 10D to 2D. The point does may be reasonable, however, Set

1, with cardinality 50 when embedded in a 2D space presents similar information preservation as the established sets from the literature, which possess on average 10 ELA metrics. Besides, this finding reveals that the "bag of metrics" approach adopted so far in the literature may not be as necessary for 2D visualization, if metrics are selected according to manifold preservation.

- Use of Co-Ranking Framework for Assessing Distortion: So far, the author hasn't seen the analysis of the impact of the embedding procedure in the lower-dimensional space from researchers in the field. In this sense, the embedded data may induce false conclusions or even present "anomalous" results as stated in (MUÑOZ; SMITH-MILES, 2017). In the present work, the Co-Ranking framework provided a suitable tool for characterizing the distortions in the manifolds by the increase or decrease of ranks between points. It made clear that a naive approach in DR could lead to erroneous findings and even provided a quantitative assessment of the degree of distortion. Besides, it opened the debate on the suitability of DR algorithms in embedding data. As shown in Chapter 4: Results and Discussions, the t-SNE and PCA presented rather distinct behavior in embedding data. The t-SNE technique produced better performance, that is, caused less distortion in the manifold, when embedding closer points (low rank). In contrast, the PCA technique produced better R_{NX} values with the increase of the considered neighborhood size, thus behaving as a better technique for preserving global distances in the studied dataset.

- Procedure for Selection of ELA metrics: It has been shown that with a singular ELA metric the classification of function categories may be difficult since many of the function categories and instances possess similar values. Therefore, the wrapper procedure of combining Genetic Algorithms with information preservation metrics has shown to be useful for producing sets of ELA metrics.

- Comparison of Established Sets of ELA Metrics: A final contribution of the present work was the comparison of different sets of ELA metrics. Some hidden patterns were repeatedly seen across the embedded sets which strengthen findings of the characteristics of the BBOB09 benchmark.

5.2 Suggestions for Future Research

To conclude, several research directions may be adopted in light of the results achieved in this work. First, one may adopt the improvements of the Co-Ranking framework. As described, the Co-Ranking in its original form only accuses the existence of distortions, it doesn't, however, assign weights to the intensity of intrusion or extrusion of points in the

lower manifold. In the same theme, other quality preservation criteria may be adopted and even compared to the Co-Ranking framework.

Second, new DR techniques may be adopted. In this work, the t-SNE and PCA presented rather distinct behavior of embedding data. However, comparing different non-linear DR may also reveal interesting information about the BBOB09 set of functions and from the performance of the algorithms themselves. In this same theme, a study on the impact of the hyperparameters of the DR algorithms in the quality of the data may also reveal interesting information to the field.

Thirdly, a suggested research direction is combining the findings of manifold preservation and classification of functions with a Machine Learning model. An interesting question to be analyzed is how does quality in manifold preservation (in spite of the embedded dimension) reflects inaccuracy in the classification model. Many Machine Learning models inherently bends the space for better classification accuracy, what are the benefits and limits of manifold preservation in the classification of unseen data?

Finally, the research of practical and real-world functions may benefit from the approach taken in this work. It produced sets of ELA metrics that may reveal interesting information about real-world functions such as seen in the fields of Engineering, Finances, etc., which may provide insight about their intrinsic nature, and ultimately in better ways for finding an optimum solution.

Bibliography

ABELL, T.; MALITSKY, Y.; TIERNEY, K. Features for exploiting black-box optimization problem structure. In: **Revised Selected Papers of the 7th International Conference on Learning and Intelligent Optimization - Volume 7997**. New York, NY, USA: Springer-Verlag New York, Inc., 2013. (LION 7), p. 30–36. ISBN 978-3-642-44972-7. Available at: <http://dx.doi.org/10.1007/978-3-642-44973-4_4>.

ALPCAN, T.; EVERITT, T.; HUTTER, M. Can we measure the difficulty of an optimization problem? In: **2014 IEEE Information Theory Workshop (ITW 2014)**, 2014. p. 356–360. ISSN 1662-9019.

BISCHL, B.; MERSMANN, O.; TRAUTMANN, H.; PREUSS, M. Algorithm selection based on exploratory landscape analysis and cost-sensitive learning. In: **Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation**. New York, NY, USA: ACM, 2012. (GECCO '12), p. 313–320. ISBN 978-1-4503-1177-9. Available at: <http://doi.acm.org/10.1145/2330163.2330209>.

BORENSTEIN, Y.; POLI, R. Kolmogorov complexity, optimization and hardness. In: **2006 IEEE International Conference on Evolutionary Computation**, 2006. p. 112–119. ISSN 1089-778X.

BOSSEK, J. smoof: Single- and multi-objective optimization test functions. **The R Journal**, 2017. Available at: <https://journal.r-project.org/archive/2017/RJ-2017-004/index.html>.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers Electrical Engineering**, v. 40, n. 1, p. 16 – 28, 2014. ISSN 0045-7906. 40th-year commemorative issue. Available at: <http://www.sciencedirect.com/science/article/pii/S0045790613003066>.

CHEN, L.; BUJA, A. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. **Journal of the American Statistical Association**, Taylor Francis, v. 104, n. 485, p. 209–219, 2009. Available at: <https://doi.org/10.1198/jasa.2009.0111>.

CLUNE, R. **Algorithm Seletion in structural optimization**. PhD Thesis (PhD Thesis) — Massachussets Institute of Technology. Department of Civil and Environmental Engineering, 2013.

COCO. **Comparing Continuous Optimizers**. 2019. Last accessed 10 February 2019. Available at: <https://coco.gforge.inria.fr/>.

CUI, C.; HU, M.; WEIR, J. D.; WU, T. A recommendation system for meta-modeling: A meta-learning based approach. **Expert Systems with Applications**, v. 46, p. 33 – 44, 2016. ISSN 0957-4174. Available at: <http://www.sciencedirect.com/science/article/pii/S0957417415007162>.

CUNNINGHAM, J. P.; GHAHRAMANI, Z. Linear dimensionality reduction: Survey, insights, and generalizations. **J. Mach. Learn. Res.**, JMLR.org, v. 16, n. 1, p. 2859–2900, jan. 2015. ISSN 1532-4435. Available at: <http://dl.acm.org/citation.cfm?id=2789272.2912091>.

GHODSI, A. **Dimensionality Reduction: A Short Tutorial**. 2006. Available at: <https://www.math.uwaterloo.ca/~aghodsib/courses/f06stat890/readings-/tutorial_stat890.pdf>.

GOGNA, A.; TAYAL, A. Metaheuristics: review and application. **Journal of Experimental & Theoretical Artificial Intelligence**, Taylor Francis, v. 25, n. 4, p. 503–526, 2013. Available at: <https://doi.org/10.1080/0952813X.2013.782347>.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization and Machine Learning**. 1st. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0201157675.

GOLLE, U. Fitness landscape analysis and design of metaheuristics for car sequencing. In: , 2011. Available at: <https://pdfs.semanticscholar.org/4e63/60330c4475c492dc84ab6b46a08e0e07f4d6.pdf>.

GONÇALVES, M. S. **Metodologia de seleção de algoritmos para problemas de otimização contínua e com formulação black-box**. Dissertation (Master of Sciences) — Universidade Federal de Santa Catarina, Brazil, 2018.

GONZÁLEZ, P.; DORTA, I.; SANTOS PEÑNATE, D.; SUÁREZ-VEGA, R. Research status and trends in operations research and management science (or/ms) journals: A bibliometric analysis based on the web of science database 2001-2012. 10 2014.

GRACIA, A.; GONZÁLEZ, S.; ROBLES, V.; MENASALVAS, E. A methodology to compare dimensionality reduction algorithms in terms of loss of quality. **Information Sciences**, v. 270, p. 1 – 27, 2014. ISSN 0020-0255. Available at: <http://www.sciencedirect.com/science/article/pii/S0020025514001741>.

HANSEN, N.; AUGER, A.; ROS, R.; FINCK, S.; POSÍK, P. Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009. In: **Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation**. New York, NY, USA: ACM, 2010. (GECCO '10), p. 1689–1696. ISBN 978-1-4503-0073-5. Available at: <http://doi.acm.org/10.1145/1830761.1830790>.

HANSEN, N.; FINCK, S.; ROS, R.; AUGER, A. **Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Functions Definitions**, 2009. Available at: <https://hal.inria.fr/inria-00362633>.

HERNANDEZ, C.; SCHÜTZE, O.; EMMERICH, M.; XIONG, F.-R.; SUN, J.-Q. Barrier tree for continuous landscapes by means of generalized cell mapping. In: **EVOLVE - A Bridge between Probability, Set Oriented Optimization, and Evolutionary Computation (Short Paper Proceedings) - July 2014, Beijing**, 2014. p. CD–ROM.

HOFFMANN, H. Kernel pca for novelty detection. **Pattern Recognition**, v. 40, n. 3, p. 863 – 874, 2007. ISSN 0031-3203. Available at: <http://www.sciencedirect.com/science/article/pii/S0031320306003414>.

HUO, X.; SMITH, A. K. A survey of manifold-based learning methods. **Recent Advances in Data Mining of Enterprise Data**, 01 2008.

JOLLIFFE, I.; CADIMA, J. Principal component analysis: A review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, p. 20150202, 04 2016.

JONES, T.; FORREST, S. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: **Proceedings of the 6th International Conference on Genetic Algorithms**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. p. 184–192. ISBN 1-55860-370-0. Available at: <http://dl.acm.org/citation.cfm?id=645514.657929>.

KERSCHKE, P. Comprehensive Feature-Based Landscape Analysis of Continuous and Constrained Optimization Problems Using the R-Package flacco. **ArXiv e-prints**, ago. 2017.

KERSCHKE, P.; PREUSS, M.; HERNÁNDEZ, C.; SCHÜTZE, O.; SUN, J.-Q.; GRIMME, C.; RUDOLPH, G.; BISCHL, B.; TRAUTMANN, H. Cell mapping techniques for exploratory landscape analysis. In: **EVOLVE - A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation V**. Cham: Springer International Publishing, 2014. p. 115–131. ISBN 978-3-319-07494-8.

KERSCHKE, P.; PREUSS, M.; WESSING, S.; TRAUTMANN, H. Detecting funnel structures by means of exploratory landscape analysis. In: **Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation**. New York, NY, USA: ACM, 2015. (GECCO '15), p. 265–272. ISBN 978-1-4503-3472-3. Available at: <http://doi.acm.org/10.1145/2739480.2754642>.

KERSCHKE, P.; TRAUTMANN, H. Automated algorithm selection on continuous black-box problems by combining exploratory landscape analysis and machine learning. **Evolutionary Computation**, v. 27, n. 1, p. 99–127, 2019. PMID: 30365386. Available at: <https://doi.org/10.1162/evco_a_00236>.

KRAEMER, G.; REICHSTEIN, M.; MAHECHA, M. dimred and coranking - unifying dimensionality reduction in r. **R Journal**, v. 10, p. 342–358, 07 2018.

LANG, M.; KOTTHAUS, H.; MARWEDEL, P.; WEIHS, C.; RAHNENFUHRER, J.; BISCHL, B. Automatic model selection for high-dimensional survival analysis. **Journal of Statistical Computation and Simulation**, Taylor Francis, v. 85, n. 1, p. 62–76, 2015. Available at: <https://doi.org/10.1080/00949655.2014.929131>.

LEARDI, R.; BOGGIA, R.; TERRILE, M. Genetic algorithms as a strategy for feature selection. **Journal of Chemometrics**, v. 6, n. 5, p. 267–281, 1992. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.1180060506>.

LEE, J. A.; VERLEYSEN, M. **Nonlinear Dimensionality Reduction**. 1st. ed.: Springer Publishing Company, Incorporated, 2007. ISBN 0387393501, 9780387393506.

LEE, J. A.; VERLEYSEN, M. Quality assessment of dimensionality reduction: Rank-based criteria. **Neurocomputing**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 72, n. 7-9, p. 1431–1443, mar. 2009. ISSN 0925-2312. Available at: <http://dx.doi.org/10.1016/j.neucom.2008.12.017>.

LEMKE, C.; BUDKA, M.; GABRYS, B. Metalearning: a survey of trends and technologies. **Artificial Intelligence Review**, v. 44, n. 1, p. 117–130, Jun 2015. ISSN 1573-7462. Available at: <https://doi.org/10.1007/s10462-013-9406-y>.

LI, B.; LI, Y.-R.; ZHANG, X.-L. A survey on laplacian eigenmaps based manifold learning methods. **Neurocomputing**, v. 335, p. 336 – 351, 2019. ISSN 0925-2312. Available at: <http://www.sciencedirect.com/science/article/pii/S0925231218312645>.

LUNACEK, M.; WHITLEY, D. The dispersion metric and the cma evolution strategy. In: **Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation**. New York, NY, USA: ACM, 2006. (GECCO '06), p. 477–484. ISBN 1-59593-186-4. Available at: <http://doi.acm.org/10.1145/1143997.1144085>.

MAATEN, L. van der; HINTON, G. Visualizing data using t-SNE. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008. Available at: <http://www.jmlr.org/papers/v9/vandermaaten08a.html>.

MALAN, K. M.; ENGELBRECHT, A. P. Quantifying ruggedness of continuous landscapes using entropy. In: **Proceedings of the Eleventh Conference on Congress on Evolutionary Computation**. Piscataway, NJ, USA: IEEE Press, 2009. (CEC'09), p. 1440–1447. ISBN 978-1-4244-2958-5. Available at: <http://dl.acm.org/citation.cfm?id=1689599.1689789>.

MALAN, K. M.; ENGELBRECHT, A. P. A survey of techniques for characterising fitness landscapes and some possible ways forward. **Information Sciences**, v. 241, p. 148 – 163, 2013. ISSN 0020-0255. Available at: <http://www.sciencedirect.com/science/article/pii/S0020025513003125>.

MCKAY, M. D. Latin hypercube sampling as a tool in uncertainty analysis of computer models. In: **Proceedings of the 24th Conference on Winter Simulation**. New York, NY, USA: ACM, 1992. (WSC '92), p. 557–564. ISBN 0-7803-0798-4. Available at: <http://doi.acm.org/10.1145/167293.167637>.

MERSMANN; TRAUTMANN; STEUER; BISCHL; DEB. **Multiple Criteria Optimization Algorithms and Related Functions**. 2014. Available at: <https://cran.r-project.org/web/packages/mco/index.html>.

MERSMANN, O.; BISCHL, B.; TRAUTMANN, H.; PREUSS, M.; WEIHS, C.; RUDOLPH, G. Exploratory landscape analysis. In: **Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation**. New York, NY, USA:

ACM, 2011. (GECCO '11), p. 829–836. ISBN 978-1-4503-0557-0. Available at: <http://doi.acm.org/10.1145/2001576.2001690>.

MERSMANN, O.; PREUSS, M.; TRAUTMANN, H. Benchmarking evolutionary algorithms: Towards exploratory landscape analysis. In: SCHAEFER, R.; COTTA, C.; KOŁODZIEJ, J.; RUDOLPH, G. (Ed.). **Parallel Problem Solving from Nature, PPSN XI**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 73–82. ISBN 978-3-642-15844-5.

MERZ, P. Advanced fitness landscape analysis and the performance of memetic algorithms. **Evol. Comput.**, MIT Press, Cambridge, MA, USA, v. 12, n. 3, p. 303–325, set. 2004. ISSN 1063-6560. Available at: <http://dx.doi.org/10.1162/1063656041774956>.

MORGAN, R.; GALLAGHER, M. Analysing and characterising optimization problems using length scale. **Soft Computing**, v. 21, n. 7, p. 1735–1752, Apr 2017. ISSN 1433-7479. Available at: <https://doi.org/10.1007/s00500-015-1878-z>.

MUÑOZ, M. A.; KIRLEY, M.; HALGAMUGE, S. K. A meta-learning prediction model of algorithm performance for continuous optimization problems. In: COELLO, C. A. C.; CUTELLO, V.; DEB, K.; FORREST, S.; NICOSIA, G.; PAVONE, M. (Ed.). **Parallel Problem Solving from Nature - PPSN XII**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 226–235. ISBN 978-3-642-32937-1.

MUÑOZ, M. A.; SMITH-MILES, K. Effects of function translation and dimensionality reduction on landscape analysis. In: **2015 IEEE Congress on Evolutionary Computation (CEC)**, 2015. p. 1336–1342. ISSN 1089-778X.

MUÑOZ, M. A. A.; SMITH-MILES, K. Performance analysis of continuous black-box optimization algorithms via footprints in instance space. **Evolutionary Computation**, v. 25, p. 529–554, 12 2017.

NOZ, M. A. M.; KIRLEY, M.; HALGAMUGE, S. Exploratory landscape analysis of continuous space optimization problems using information content. v. 19, p. 74–87, 01 2015.

PLESS, R.; SOUVENIR, R. A survey of manifold learning for images. **Ipsj Transactions on Computer Vision and Applications**, v. 1, p. 83–94, 01 2009.

PUNCH, W.; GOODMAN, E.; PEI, M.; CHIA-SHUN, L.; HOVLAND, P. D.; ENBODY, R. Further research on feature selection and classification using genetic algorithms. In: , 1993. p. 557–564.

QU, B.; LIANG, J.; WANG, Z.; CHEN, Q.; SUGANTHAN, P. Novel benchmark functions for continuous multimodal optimization with comparative results. **Swarm and Evolutionary Computation**, v. 26, p. 23 – 34, 2016. ISSN 2210-6502. Available at: <http://www.sciencedirect.com/science/article/pii/S221065021500053X>.

RICE, J. R. The algorithm selection problem**this work was partially supported by the national science foundation through grant gp-32940x. this chapter was presented as the george e. forsythe memorial lecture at the computer science conference, february 19, 1975, washington, d. c. In: RUBINOFF, M.; YOVITS, M. C. (Ed.). Elsevier, 1976,

(Advances in Computers, v. 15). p. 65 – 118. Available at: <http://www.sciencedirect.com/science/article/pii/S0065245808605203>.

SILVA, V. D.; TENENBAUM, J. B. Global versus local methods in nonlinear dimensionality reduction. In: **Advances in Neural Information Processing Systems 15**: MIT Press, 2003. p. 705–712.

SMITH-MILES, K. A. Cross-disciplinary perspectives on meta-learning for algorithm selection. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 41, n. 1, p. 6:1–6:25, jan. 2009. ISSN 0360-0300. Available at: <http://doi.acm.org/10.1145/1456650.1456656>.

STADLER, P. F. Landscapes and their correlation functions. **Journal of Mathematical Chemistry**, v. 20, n. 1, p. 1–45, Mar 1996. ISSN 1572-8897. Available at: <https://doi.org/10.1007/BF01165154>.

SUGANTHAN; HANSEN; LIANG; DEB; CHEN; AUGER; TIWARI. **Problem Definitions and Evaluation Criteria for the CEC 2005 Special Session on Real-Parameter Optimization**, 2005. Available at: <http://www.ntu.edu.sg/home/EPNSugan/index_files/CEC-05/CEC05.htm>. Access in: 12 dec. 2016.

SUN, Y.; HALGAMUGE, S. K.; KIRLEY, M.; MUNOZ, M. A. On the selection of fitness landscape analysis metrics for continuous optimization problems. In: **7th International Conference on Information and Automation for Sustainability**, 2014. p. 1–6. ISSN 2151-1802.

VASSILEV, V. K.; FOGARTY, T. C.; MILLER, J. F. Information characteristics and the structure of landscapes. **Evol. Comput.**, MIT Press, Cambridge, MA, USA, v. 8, n. 1, p. 31–60, mar. 2000. ISSN 1063-6560. Available at: <http://dx.doi.org/10.1162/106365600568095>.

WATSON, J.-P. **An Introduction to Fitness Landscape Analysis and Cost Models for Local Search**. Boston, MA: Springer US, 2010. 599–623 p. ISBN 978-1-4419-1665-5. Available at: <https://doi.org/10.1007/978-1-4419-1665-5_20>.

WEISE, T. **Global Optimization Algorithms - Theory and Application**. Second. Self-Published, 2009. Online as e-book. Online available at http://www.it-weise.de/. Available at: <http://www.it-weise.de/>.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. **Trans. Evol. Comp**, IEEE Press, Piscataway, NJ, USA, v. 1, n. 1, p. 67–82, abr. 1997. ISSN 1089-778X. Available at: <https://doi.org/10.1109/4235.585893>.

YANG, J.; HONAVAR, V. **Feature Subset Selection Using a Genetic Algorithm**. Boston, MA: Springer US, 1998. 117–136 p. ISBN 978-1-4615-5725-8. Available at: <https://doi.org/10.1007/978-1-4615-5725-8_8>.

ZHENG, N.; XUE, J. **Manifold Learning**. London: Springer London, 2009. 87–119 p. ISBN 978-1-84882-312-9. Available at: <https://doi.org/10.1007/978-1-84882-312-9_4>.

ZITZLER, E.; DEB, K.; THIELE, L. Comparison of multiobjective evolutionary algorithms: Empirical results. **Evol. Comput.**, MIT Press, Cambridge, MA, USA, v. 8, n. 2, p. 173–195, jun. 2000. ISSN 1063-6560. Available at: <http://dx.doi.org/10.1162/106365600568202>.

Appendix A - Definition of the Black Box Optimization Benchmark Functions

From (HANSEN *et al.*, 2009)

TABLE A.1 – Properties of the BBOB09 functions

Function Type	**Properties**	**Category**
Sphere	Unimodal Highly Symmetric	Separable
Ellipsoidal	Unimodal Conditioning about 10ˆ6	Separable
Rastrigin	Roughly 10ˆD local optima Conditioning about 10	Separable
Büche-Rastringin	Roughly 10ˆD local optima Conditioning about 10 Skewed in x and f-space	Separable
Linear Slope	Optimum is on domain boundary	Separable
Attractive Sector	Unimodal Highly assymetric	Low or moderate conditioning
Step Ellipsoidal	Consists of many plateaus of different sizes Conditioning about 10ˆ2	Low or moderate conditioning
Rosenbrock	It possesses different attraction basins according to dimensionality of the problem	Low or moderate conditioning
Rosenbrock Rotated	It possesses different attraction basins according to dimensionality of the problem	Low or moderate conditioning
Ellipsoidal	Ill condionted Smooth local irregularities Conditioning about 10ˆ6	High conditioning and unimodal

TABLE A.1 – Properties of the BBOB09 functions - Cont.

Discus	Globally quadratic Local irregularities Conditioning about 10ˆ6	High conditioning and unimodal
Bent Cigar	Smoot and narrow ridges Overall shape deviates from quadratic Conditioning about 10ˆ6	High conditioning and unimodal
Sharp Ridge	Slope 1:100 Neutral area	High conditioning and unimodal
Different Powers	The sensitivity relation of the variables worsens when close to the optimum	High conditioning and unimodal
Rastrigin	Non-separable Roughly 10ˆD local optima Conditining about 10	Multi-modal with adequate global structure
Weierstrass	Highly rugged Moderately repetitive	Multi-modal with adequate global structure
Schaffers F7	Highly multimodal with varying frequency and amplitude of modulation Low conditioning	Multi-modal with adequate global structure
Schaffers F7 Moderately Ill-Condtioned	Moderately ill-conditoned counter part of f17 Conditioning about 10ˆ3	Multi-modal with adequate global structure
Composite Griewank-Rosenbrock	Highly multimodal	Multi-modal with adequate global structure
Schwefel	Landscape with penalized and unpennalized area	Multi-modal with weak global structure
Gallaghers Gaussian 101-me Peaks	101 optima with different positioning Conditioning about 30 around global optimum	Multi-modal with weak global structure
Gallaghers Gaussian 21-hi Peaks	21 optima with different positioning Conditioning about 10ˆ3 around global optimum	Multi-modal with weak global structure
Katsura	Highly rugged Highly repetitive More than 10ˆD global optima	Multi-modal with weak global structure
Lunacek bi-Rastrigin	Highly multimodal with the presence of funnels	Multi-modal with weak global structure

1. Sphere Function

$$f_1(\mathbf{x}) = \|\mathbf{z}\|^2 + f_{\text{opt}}$$

- $\mathbf{z} = \mathbf{x} - \mathbf{x}^{\text{opt}}$

2. Ellipsoidal Function

$$f_2(\mathbf{x}) = \sum_{i=1}^{D} 10^{6\frac{i-1}{D-1}} z_i^2 + f_{\text{opt}}$$

- $\mathbf{z} = T_{\text{osz}}(\mathbf{x} - \mathbf{x}^{\text{opt}})$

3. Rastrigin Function

$$f_3(\mathbf{x}) = 10\left(D - \sum_{i=1}^{D}\cos(2\pi z_i)\right) + \|\mathbf{z}\|^2 + f_{\text{opt}}$$

- $\mathbf{z} = \Lambda^{10} T_{\text{asy}}^{0.2}(T_{\text{osz}}(\mathbf{x} - \mathbf{x}^{\text{opt}}))$

4. Büche-Rastrigin Function

$$f_4(\mathbf{x}) = 10\left(D - \sum_{i=1}^{D}\cos(2\pi z_i)\right) + \sum_{i=1}^{D} z_i^2 + 100\, f_{\text{pen}}(\mathbf{x}) + f_{\text{opt}}$$

- $z_i = s_i\, T_{\text{osz}}(x_i - x_i^{\text{opt}}) \quad \text{for } i = 1 \ldots D$
- $s_i = \begin{cases} 10 \times 10^{\frac{1}{2}\frac{i-1}{D-1}} & \text{if } z_i > 0 \text{ and } i = 1, 3, 5, \ldots \\ 10^{\frac{1}{2}\frac{i-1}{D-1}} & \text{otherwise} \end{cases} \quad \text{for } i = 1, \ldots, D$

5. Linear Slope

$$f_5(\mathbf{x}) = \sum_{i=1}^{D} 5\,|s_i| - s_i z_i + f_{\text{opt}} \tag{5}$$

- $z_i = x_i$ if $x_i^{\text{opt}} x_i < 5^2$ and $z_i = x_i^{\text{opt}}$ otherwise, for $i = 1, \ldots, D$. That is, if x_i exceeds x_i^{opt} it will mapped back into the domain and the function appears to be constant in this direction.
- $s_i = \text{sign}\left(x_i^{\text{opt}}\right) 10^{\frac{i-1}{D-1}}$ for $i = 1, \ldots, D$.
- $\mathbf{x}^{\text{opt}} = \mathbf{z}^{\text{opt}} = 5 \times \mathbf{1}_{-}^{+}$

6. Attractive Sector Function

$$f_6(\mathbf{x}) = T_{\text{osz}}\left(\sum_{i=1}^{D}(s_i z_i)^2\right)^{0.9} + f_{\text{opt}}$$

- $\mathbf{z} = \mathbf{Q}\Lambda^{10}\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}})$
- $s_i = \begin{cases} 10^2 & \text{if } z_i > 0 \\ 1 & \text{otherwise} \end{cases}$

7. Step Ellipsoidal Function

$$f_7(\mathbf{x}) = 0.1 \max\left(|\hat{z}_1|/10^4, \sum_{i=1}^{D} 10^{2\frac{i-1}{D-1}} z_i^2\right) + f_{\text{pen}}(\mathbf{x}) + f_{\text{opt}}$$

- $\hat{\mathbf{z}} = \Lambda^{10}\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}})$
- $\tilde{z}_i = \begin{cases} \lfloor 0.5 + \hat{z}_i \rfloor & \text{if } \hat{z}_i > 0.5 \\ \lfloor 0.5 + 10\,\hat{z}_i \rfloor / 10 & \text{otherwise} \end{cases} \text{ for } i = 1, \ldots, D,$
  denotes the rounding procedure in order to produce the plateaus.
- $\mathbf{z} = \mathbf{Q}\tilde{\mathbf{z}}$

8. Rosenbrock Function

$$f_8(\mathbf{x}) = \sum_{i=1}^{D-1} \left(100\ \left(z_i^2 - z_{i+1}\right)^2 + (z_i - 1)^2\right) + f_{\text{opt}}$$

- $\mathbf{z} = \max\left(1, \frac{\sqrt{D}}{8}\right)(\mathbf{x} - \mathbf{x}^{\text{opt}}) + 1$
- $\mathbf{z}^{\text{opt}} = \mathbf{1}$

9. Rosenbrock Function, Rotated

$$f_9(\mathbf{x}) = \sum_{i=1}^{D-1} \left(100\ \left(z_i^2 - z_{i+1}\right)^2 + (z_i - 1)^2\right) + f_{\text{opt}}$$

- $\mathbf{z} = \max\left(1, \frac{\sqrt{D}}{8}\right)\mathbf{R}\mathbf{x} + \mathbf{1}/2$
- $\mathbf{z}^{\text{opt}} = \mathbf{1}$

10. Ellipsoidal Function

$$f_{10}(\mathbf{x}) = \sum_{i=1}^{D} 10^{6\frac{i-1}{D-1}} z_i^2 + f_{\text{opt}}$$

- $\mathbf{z} = T_{\text{osz}}(\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}}))$

11. Discus Function

$$f_{11}(\mathbf{x}) = 10^6 z_1^2 + \sum_{i=2}^{D} z_i^2 + f_{\text{opt}}$$

- $\mathbf{z} = T_{\text{osz}}(\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}}))$

12. Bent Cigar Function

$$f_{12}(\mathbf{x}) = z_1^2 + 10^6 \sum_{i=2}^{D} z_i^2 + f_{\text{opt}}$$

- $\mathbf{z} = \mathbf{R}\, T_{\text{asy}}^{0.5}(\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}}))$

13. Sharp Ridge Function

$$f_{13}(\mathbf{x}) = z_1^2 + 100\sqrt{\sum_{i=2}^{D} z_i^2} + f_{\text{opt}}$$

- $\mathbf{z} = \mathbf{Q}\Lambda^{10}\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}})$

14. Different Powers Function

$$f_{14}(\mathbf{x}) = \sqrt{\sum_{i=1}^{D} |z_i|^{2+4\frac{i-1}{D-1}}} + f_{\text{opt}}$$

- $\mathbf{z} = \mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}})$

15. Rastrigin Function

$$f_{15}(\mathbf{x}) = 10\left(D - \sum_{i=1}^{D} \cos(2\pi z_i)\right) + \|\mathbf{z}\|^2 + f_{\text{opt}}$$

- $\mathbf{z} = \mathbf{R}\Lambda^{10}\mathbf{Q}\, T_{\text{asy}}^{0.2}(T_{\text{osz}}(\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}})))$

16. Weierstrass Function

$$f_{16}(\mathbf{x}) = 10\left(\frac{1}{D}\sum_{i=1}^{D}\sum_{k=0}^{11} 1/2^k \cos(2\pi 3^k(z_i + 1/2)) - f_0\right)^3 + \frac{10}{D} f_{\text{pen}}(\mathbf{x}) + f_{\text{opt}}$$

- $\mathbf{z} = \mathbf{R}\Lambda^{1/100}\mathbf{Q}\, T_{\text{osz}}(\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}}))$
- $f_0 = \sum_{k=0}^{11} 1/2^k \cos(2\pi 3^k 1/2)$

17. Schaffers F7 Function

$$f_{17}(\mathbf{x}) = \left(\frac{1}{D-1} \sum_{i=1}^{D-1} \sqrt{s_i} + \sqrt{s_i} \sin^2\left(50\, s_i^{1/5}\right) \right)^2 + 10\, f_{\text{pen}}(\mathbf{x}) + f_{\text{opt}}$$

- $\mathbf{z} = \Lambda^{10} \mathbf{Q}\, T_{\text{asy}}^{0.5}(\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}}))$
- $s_i = \sqrt{z_i^2 + z_{i+1}^2}$ for $i = 1, \ldots, D$

18. Schaffers F7 Function, moderately ill-conditioned

$$f_{18}(\mathbf{x}) = \left(\frac{1}{D-1} \sum_{i=1}^{D-1} \sqrt{s_i} + \sqrt{s_i} \sin^2\left(50\, s_i^{1/5}\right) \right)^2 + 10\, f_{\text{pen}}(\mathbf{x}) + f_{\text{opt}}$$

- $\mathbf{z} = \Lambda^{1000} \mathbf{Q}\, T_{\text{asy}}^{0.5}(\mathbf{R}(\mathbf{x} - \mathbf{x}^{\text{opt}}))$
- $s_i = \sqrt{z_i^2 + z_{i+1}^2}$ for $i = 1, \ldots, D$

19. Composite Griewank-Rosenbrock Function F8F2

$$f_{19}(\mathbf{x}) = \frac{10}{D-1} \sum_{i=1}^{D-1} \left(\frac{s_i}{4000} - \cos(s_i) \right) + 10 + f_{\text{opt}}$$

- $\mathbf{z} = \max\left(1, \frac{\sqrt{D}}{8}\right) \mathbf{R}\mathbf{x} + 0.5$
- $s_i = 100\,(z_i^2 - z_{i+1})^2 + (z_i - 1)^2$ for $i = 1, \ldots, D$
- $\mathbf{z}^{\text{opt}} = \mathbf{1}$

20. Schwefel Function

$$f_{20}(\mathbf{x}) = -\frac{1}{D} \sum_{i=1}^{D} z_i \sin(\sqrt{|z_i|}) + 4.189828872724339 + 100 f_{\text{pen}}(\mathbf{z}/100) + f_{\text{opt}}$$

- $\hat{\mathbf{x}} = 2 \times \mathbf{1}_-^+ \otimes \mathbf{x}$
- $\hat{z}_1 = \hat{x}_1$, $\hat{z}_{i+1} = \hat{x}_{i+1} + 0.25\,(\hat{x}_i - x_i^{\text{opt}})$ for $i = 1, \ldots, D-1$
- $\mathbf{z} = 100\,(\Lambda^{10}(\hat{\mathbf{z}} - \mathbf{x}^{\text{opt}}) + \mathbf{x}^{\text{opt}})$
- $\mathbf{x}^{\text{opt}} = 4.2096874633/2\ \mathbf{1}_-^+$, where $\mathbf{1}_-^+$ is the same realization as above

21. Gallagher's Gaussian 101-me Peaks Function

$$f_{21}(\mathbf{x}) = T_{\text{osz}}\left(10 - \max_{i=1}^{101} w_i \exp\left(-\frac{1}{2D} (\mathbf{x} - \mathbf{y}_i)^{\mathrm{T}} \mathbf{R}^{\mathrm{T}} \mathbf{C}_i \mathbf{R} (\mathbf{x} - \mathbf{y}_i) \right) \right)^2 + f_{\text{pen}}(\mathbf{x}) + f_{\text{opt}}$$

- $w_i = \begin{cases} 1.1 + 8 \times \frac{i-2}{99} & \text{for } i = 2, \ldots, 101 \\ 10 & \text{for } i = 1 \end{cases}$, three optima have a value larger than 9
- $\mathbf{C}_i = \Lambda^{\alpha_i} / \alpha_i^{1/4}$ where Λ^{α_i} is defined as usual (see Section 0.2), but with randomly permuted diagonal elements. For $i = 2, \ldots, 101$, α_i is drawn uniformly randomly from the set $\left\{ 1000^{2\frac{j}{99}} \mid j = 0, \ldots, 99 \right\}$ without replacement, and $\alpha_i = 1000$ for $i = 1$.
- the local optima $\mathbf{y}_i$ are uniformly drawn from the domain $[-4.9, 4.9]^D$ for $i = 2, \ldots, 101$ and $\mathbf{y}_1 \in [-4, 4]^D$. The global optimum is at $\mathbf{x}^{\text{opt}} = \mathbf{y}_1$.

22. Gallagher's Gaussian 21-hi Peaks Function

$$f_{22}(\mathbf{x}) = T_{\text{osz}}\left(10 - \max_{i=1}^{21} w_i \exp\left(-\frac{1}{2D} (\mathbf{x} - \mathbf{y}_i)^{\mathrm{T}} \mathbf{R}^{\mathrm{T}} \mathbf{C}_i \mathbf{R} (\mathbf{x} - \mathbf{y}_i) \right) \right)^2 + f_{\text{pen}}(\mathbf{x}) + f_{\text{opt}}$$

- $w_i = \begin{cases} 1.1 + 8 \times \frac{i-2}{19} & \text{for } i = 2, \ldots, 21 \\ 10 & \text{for } i = 1 \end{cases}$, two optima have a value larger than 9
- $\mathbf{C}_i = \Lambda^{\alpha_i} / \alpha_i^{1/4}$ where Λ^{α_i} is defined as usual (see Section 0.2), but with randomly permuted diagonal elements. For $i = 2, \ldots, 21$, α_i is drawn uniformly randomly from the set $\left\{ 1000^{2\frac{j}{19}} \mid j = 0, \ldots, 19 \right\}$ without replacement, and $\alpha_i = 1000^2$ for $i = 1$.
- the local optima $\mathbf{y}_i$ are uniformly drawn from the domain $[-4.9, 4.9]^D$ for $i = 2, \ldots, 21$ and $\mathbf{y}_1 \in [-4, 4]^D$. The global optimum is at $\mathbf{x}^{\text{opt}} = \mathbf{y}_1$.

23. Katsura Function

$$f_{23}(\mathbf{x}) = \frac{10}{D^2} \prod_{i=1}^{D} \left(1 + i \sum_{j=1}^{32} \frac{\left| 2^j z_i - [2^j z_i] \right|}{2^j} \right)^{10/D^{1.2}} - \frac{10}{D^2} + f_{\text{pen}}(\mathbf{x})$$

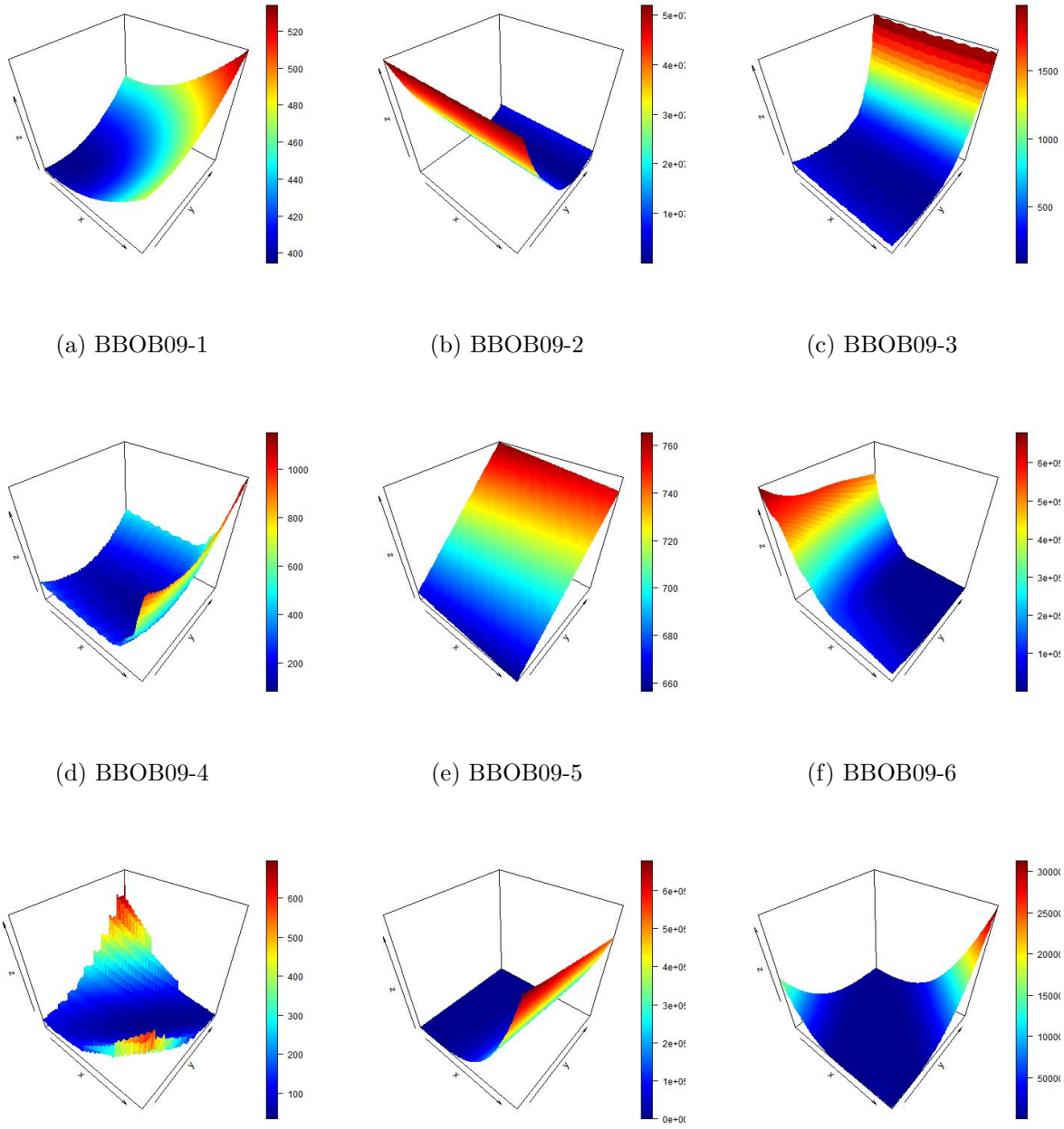
- $\mathbf{z} = \mathbf{Q}\, \Lambda^{100} \mathbf{R} (\mathbf{x} - \mathbf{x}^{\text{opt}})$

24. Lunacek Function

$$f_{24}(\mathbf{x}) = \min \left(\sum_{i=1}^{D} (\hat{x}_i - \mu_0)^2, d\, D + s \sum_{i=1}^{D} (\hat{x}_i - \mu_1)^2 \right) + 10 \left(D - \sum_{i=1}^{D} \cos(2\pi z_i) \right) + 10^4 f_{\text{pen}}(\mathbf{x}) \tag{24}$$

- $\hat{\mathbf{x}} = 2 \operatorname{sign}(\mathbf{x}^{\text{opt}}) \otimes \mathbf{x}$
- $\mathbf{z} = \mathbf{Q} \Lambda^{100} \mathbf{R} (\hat{\mathbf{x}} - \mu_0\, \mathbf{1})$
- $\mu_0 = 2.5,\ \mu_1 = -\sqrt{\frac{\mu_0^2 - d}{s}},\ s = 1 - \frac{1}{2\sqrt{D+20} - 8.2},\ d = 1$
- $\mathbf{x}^{\text{opt}} = \mu_0 \mathbf{1}^+_-$

Appendix B - Landscape of the BBOB09 Functions

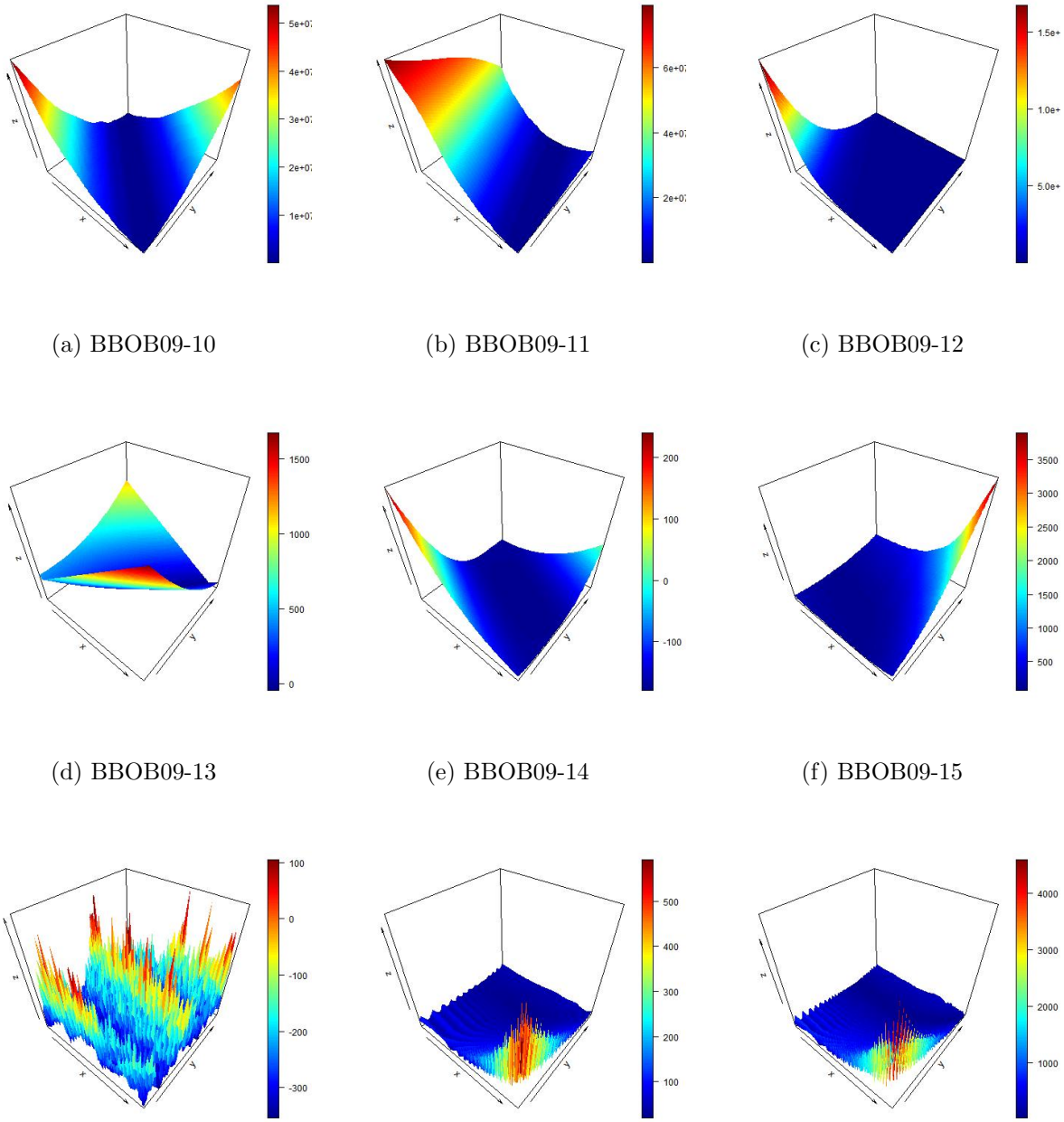


(a) BBOB09-1 (b) BBOB09-2 (c) BBOB09-3

(d) BBOB09-4 (e) BBOB09-5 (f) BBOB09-6

(g) BBOB09-7 (h) BBOB09-8 (i) BBOB09-9

FIGURE B.1 – BBOB Functions 1 - 9.

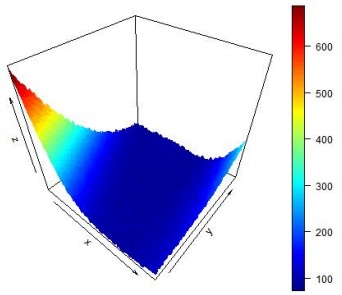


(a) BBOB09-10 (b) BBOB09-11 (c) BBOB09-12

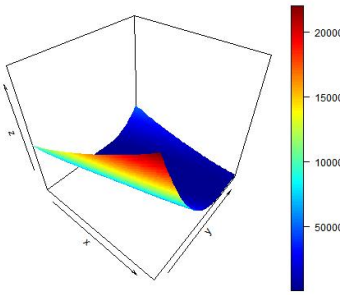
(d) BBOB09-13 (e) BBOB09-14 (f) BBOB09-15

(g) BBOB09-16 (h) BBOB09-17 (i) BBOB09-18

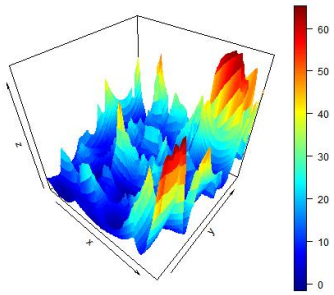
FIGURE B.2 – BBOB Functions 10 - 18.



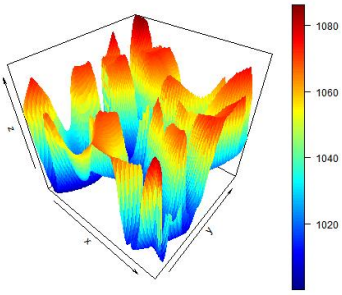
(a) BBOB09-19



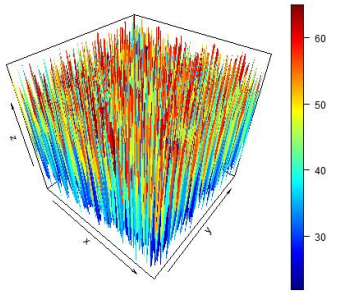
(b) BBOB09-20



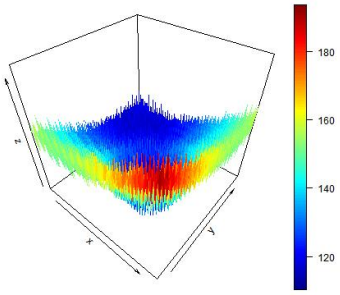
(c) BBOB09-21



(d) BBOB09-22



(e) BBOB09-23



(f) BBOB09-24

FIGURE B.3 – BBOB Functions 19 - 24.

Appendix C - Description of the Adopted ELA Metrics

In this section, the definitions of the adopted ELA metrics are provided.

C.1 Fitness Distance Correlation

The Fitness Distance Correlation as defined in (JONES; FORREST, 1995) measures function Ruggedness, Neutrality and Multi-Modality characteristics. (GONÇALVES, 2018)

$$FDC = \frac{C_{FD}}{\sigma_f \sigma_d} \tag{C.1}$$

$$C_{FD} = \frac{1}{n} \sum_{i=1}^{i=n} (f_i - f_*)(d_i - d_*) \tag{C.2}$$

Where d_* and f_* denote the mean distance of all points to the coordinates of a defined optimum and the optimal point fitness value, respectively. If the defined optimum is the global known optimal, the FDC_{global} is calculated. Conversely, if * is defined as the local optimum from a sample, then FDC_{local} is calculated.

C.2 Correlation Length

The Correlation Length (l) (STADLER, 1996) is based on the principle of random walk and autocorrelation. As described in (GOLLE, 2011), beginning with an arbitrary data point, a random walk picks a random solution in the neighborhood of the current solution and proceeds the walk with the new solution. The procedure is repeated m times. Similar

to the FDC metric, the Correlation Lenght measures a function ruggedness.

$$r(s) = \frac{1}{\sigma^2(f)(m-s)} \sum_{t=1}^{m-s} (f(x_t) - \bar{f})(f(x_{t+s}) - \bar{f}) \tag{C.3}$$

$$l = \begin{cases} 0, & \text{if } r(l) = 0 \\ -\frac{1}{ln(|r(l)|)}, & \text{otherwise} \end{cases} \tag{C.4}$$

C.3 Information Contents

Defined in (VASSILEV *et al.*, 2000), the Information Content based metrics were developed to capture the amount of information embedded in a function landscape. Further, the authors continue that from a sequence of fitness values $f_{t\,t=0}^{n}$ a discretization to a string of n symbols $S(\epsilon) = s1s2s3...sn$ of symbols $s_i \in [\bar{1}, 0, 1]$ is performed as follows:

$$s_i = \psi_{f_t}(i, \epsilon) \tag{C.5}$$

$$\psi_{f_t}(i, \epsilon) = \begin{cases} \bar{1}, & \text{if } f_i - f_{i-1} < -\epsilon \\ 0, & \text{if } |f_i - f_{i-1}| \leq \epsilon \\ 1, & \text{if } f_i - f_{i-1} > \epsilon \end{cases} \tag{C.6}$$

Then the metrics are defined as in (VASSILEV *et al.*, 2000):

- Information content: An entropic measure of the fitness sequence.
- Partial information content: A metric of the modality of the landscape path in the fitness sequence.
- Information stability: Measures the highest fitness difference between neighboring points in the fitness sequence.

C.3.1 Information Content

The entropic measure is defined as:

$$H(\epsilon) = -\sum_{p \neq q} P_{[pq]} log_6 P_{[pq]} \tag{C.7}$$

$P_{[pq]}$ is the number of occurrences of in a given block of symbols pq from the set $\bar{1}, 0, 1$.

$$P_{[pq]} = \frac{n_{[pq]}}{n} \tag{C.8}$$

C.3.2 Partial Information Content

A new string is defined based on $S(\epsilon)$ as follows: $S'(\epsilon)$ is empty if $S(\epsilon)$ is a sequence of 0s, otherwise $S'(\epsilon) = s1s2...sn$. $S'(\epsilon)$ possesses length μ. Therefore, the Partial Information Content is defined as:

$$M(\epsilon) = \frac{\mu}{n} \tag{C.9}$$

C.3.3 Information Stability

Since the magnitude of ϵ directly affects the variety of symbols defined in the Information Content metric, the authors define a new metric called Information Stability. The Information Stability is the minimum value of ϵ in which the sequences $S(\epsilon)$ and $S'(\epsilon)$ become flat.

Also, in (KERSCHKE, 2017) the authors introduce new metrics based on the statistics of the information metrics calculated above. For instance, adopted metrics in this work:

- •Max Entropy: Maximum entropy according to the variation of ϵ.
- •Max ϵ: Medium values of ϵ which provides maximum entropy value.
- •ϵ Settling Sensitivity: Subsets all entropy values H below a threshold. Pick the minimum value of this subset and calculate the final metric such as $log_{10}min(H)$.
- •ϵ ratio: Subsets all entropy values M above a threshold. Pick the maximum value of this subset and calculate the final metric such as $log_{10}max(M)$.

C.4 Length Scale

The Length Scale metric was proposed in (MORGAN; GALLAGHER, 2017) and, given two point x_1 and x_2, it is defined as follows:

$$LS(x_1, x_2) = \frac{|f(x_1) - f(x_2)|}{||x_1 - x_2||} \tag{C.10}$$

Therefore, the Length Scale metric is calculated using all pairs of points. Statistics from the results are calculated. In the present work the following statistics are adopted: Shannon Entropy, Maximum, Minimum, Variance, Mode, Mean and Kernel Bandwidth.

C.5 Fitness (Y) Distribution

In this family of ELA metrics, the calculated values are based directly on the fitness values from the sample points. The adopted statistics are Skewness, Kurtosis, and Number of Peaks in the histogram.

C.6 Dispersion

The Dispersion family of metrics, as defined in (LUNACEK; WHITLEY, 2006), is the average distance between the best points of a sample. The best points are defined as the $\alpha\%$ points that possess the lowest fitness values (in the case of a minimization problem).

$$D = \frac{1}{(\alpha Np)(\alpha Np - 1)} \sum_{i=1}^{\alpha Np} \sum_{j=1, j \neq i}^{\alpha Np} d(x_i, x_j) \tag{C.11}$$

Also, because this metric naturally separates points in groups belonging to top-ranked and non-top ranked, other ELA metrics may be derived as shown in (KERSCHKE, 2017). The adopted metrics are:

- •Ratio Statistics: Statistics from the top-ranked points are divided by the same statistics from the whole sample. For e.g: $Dmean = \frac{meanTop5\%}{meanTotalSample}$
- •Difference Statistics: Statistics from the top-ranked points are subtracted by the same statistics from the whole sample. For e.g: $Dmean = meanTop5\% - meanTotalSample$

C.7 Local Search

This family of ELA metrics, as defined in (MERSMANN *et al.*, 2010), employ a local search optimization algorithm (Nelder Mead). Starting with a random subsample of size $N = 50d$ from the original sample, the solutions of the search are hierarchically clustered to identify the local optima. From such clustering, several metrics are extracted as shown in the list:

- Statistics in the Function Evaluations: Statistics are calculated in the new subsample, such as Minimum, Lower Quartile, Mean, Median, Upper Quartile, Maximum and Standard Deviation.

- Number of Clusters: Absolute number of clusters of local optima found in the Hierarchical Clustering, as well as its relative number, that is, the number of clusters divided by the number of points in the sample.

- Contrast Analysis: Statistics such as the best fitness value of a solution divided by the mean fitness values of the local optima.

- Cluster Size: Statistics of the size of clusters of local optima such a Mean and Median.

C.8 Meta-Models

In the Meta-Models family of ELA metrics, linear and quadratic regressions are fitted in the sample data(MERSMANN *et al.*, 2010). The R^2_{adj} coefficient is also calculated as a metric for the model adequacy. The ELA metrics based on these family of models are defined as follows:

In a 2D example, the regression models would be defined as follows:

Linear Model without Interaction

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{C.12}$$

Linear Model with Interaction

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \tag{C.13}$$

Quadratic Model without Interaction

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon \tag{C.14}$$

Quadratic Model with Interaction

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \tag{C.15}$$

The ELA metrics based on these family of models are then defined as follows:

- R^2_{adj} Pearson Coefficient: The meta-model accuracy of both linear and quadratic models are used as ELA metrics.
- Model Intercept: Model intercept β_0 for the linear models with and without interaction.
- Statistics on Intercept: Minimum, maximum and max/min ratio of the absolute values of the model coefficients β_i.
- Model Condition: The ratio between the maximum absolute quadratic β_{ik} coefficient to the minimum absolute quadratic coefficient β_{ij}.

C.9 Convexity

This family of ELA metrics measures the convexity and the probability of the convexity within a specified sample. As defined in (MERSMANN *et al.*, 2010), two random points from the sample are selected and a linear combination with random weights is formed. The difference between the fitness value of the new point and the original fitness is calculated (δ). The procedure, as defined in the original paper, is replicated 1000 times.

Then, statistics from the replicated results are calculated:

- Convexity Probability: Mean number of times the difference δ is below a defined threshold ϵ. In (MERSMANN *et al.*, 2010) and (KERSCHKE, 2017), $\epsilon = 10^{-10}$.
- Linear Probability: Mean number of times the absolute of the difference δ is less than ϵ.
- Linear Deviation: Mean value of the difference δ values.
- Absolute Linear Deviation: Mean value of the absolutes of the differences δ values.

C.10 Level Set

In this family of ELA metrics defined in (MERSMANN *et al.*, 2010), a sample is divided according to a pre-defined quantile α of its fitness values. The divided groups serve as input for three different classification models, namely Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Mixture Discriminant Analysis (MDA). Their median misclassification errors (MMCE) are adopted as ELA metrics for estimating characteristics of the sample. Two types of metrics are defined as follows:

•Ratio of Errors: It calculates the ratio of the misclassification errors from two discriminant techniques for a given α.

•Median Misclassification Error (MMCE): It calculates the misclassifiation error from a discrimant technique for a given α.

C.11 Curvature

The Curvature family of ELA metrics use statistics of the numerically calculated gradients in each point of a subsample of size $s = 100d$ (MERSMANN *et al.*, 2010). Some defined metrics are:

•Statistics of the Normalized Sample of Gradients: Each value in the sample of calculated gradients is normalized by the square root of the squared sum of the sample. With this new, normalized sample, statistics are calculated. The adopted statistics are Minimum, Lower Quartile, Mean, Median, Upper Quartile and Maximum.

•Statistics of the Scaled Sample of Gradients: Each value in the sample of calculated gradients is scaled by a constant ratio, defined as the maximum gradient divided by the minimum gradient value in the sample. With this new, scaled sample, statistics are calculated. The adopted statistics are Minimum, Lower Quartile, Mean, Median, Upper Quartile and Maximum.

•Statistics of the Condition Number of the Hessian Matrices: The condition number of each Hessian Matrix defined by the sample is calculated as follows: The ratio of the maximum eigenvalue by the minimum eigenvalue. With the sample of condition numbers, statistics are calculated. The adopted statistics are Minimum, Lower Quartile, Mean, Median, Upper Quartile and Maximum.

Appendix D - Table of Selected ELA Metrics - Superset

	Feature Name	**Flacco Identification**	**Low Level Property/Family**
1	Linear Probability	ela_conv.lin_prob	Convexity
2	Linear Deviation	ela_conv.lin_dev.orig	Convexity
3	Absolute Linear Deviation	ela_conv.lin_dev.abs	Convexity
4	Minimum of Normalized Gradient	ela_curv.grad_norm.min	Curvature
5	Lower Quartile of Normalized Gradient	ela_curv.grad_norm.lq	Curvature
6	Mean Normalized Gradient	ela_curv.grad_norm.mean	Curvature
7	Median Normalized Gradient	ela_curv.grad_norm.med	Curvature
8	Upper Quartile of Normalized Gradient	ela_curv.grad_norm.uq	Curvature
9	Maximum of Normalized Gradient	ela_curv.grad_norm.max	Curvature
10	Std. Dev Normalized Gradient	ela_curv.grad_norm.sd	Curvature
11	Number of Nas Normalized Gradient	ela_curv.grad_norm.nas	Curvature
12	Lower Quartile of Scaled Gradient	ela_curv.grad_scale.min	Curvature

13	Mean of Scaled Gradient	ela_curv.grad_scale.lq	Curvature
14	Median of Scaled Gradient	ela_curv.grad_scale.mean	Curvature
15	Upper Quartile of Scaled Gradient	ela_curv.grad_scale.med	Curvature
16	Maximum of Scaled Gradient	ela_curv.grad_scale.uq	Curvature
17	Maximum Scaled Gradient	ela_curv.grad_scale.max	Curvature
18	Std. Dev Scaled Gradient	ela_curv.grad_scale.sd	Curvature
19	Number of Nas Scaled Gradient	ela_curv.grad_scale.nas	Curvature
20	Condition Number of Hessian Matrix Minimum	ela_curv.hessian_cond.min	Curvature
21	Condition Number of Hessian Matrix Lower Quartile	ela_curv.hessian_cond.lq	Curvature
22	Condition Number of Hessian Matrix Mean	ela_curv.hessian_cond.mean	Curvature
23	Condition Number of Hessian Matrix Lower Median	ela_curv.hessian_cond.med	Curvature
24	Condition Number of Hessian Matrix Upper Quartile	ela_curv.hessian_cond.uq	Curvature
25	Condition Number of Hessian Matrix Maximum	ela_curv.hessian_cond.max	Curvature
26	Condition Number of Hessian Matrix Std. Dev	ela_curv.hessian_cond.sd	Curvature
27	Condition Number of Hessian Matrix N Nas	ela_curv.hessian_cond.nas	Curvature
28	Fitness Distribution Skewness	ela_distr.skewness	Y Probability Distribution

29	Fitness Distribution Kurtosis	ela_distr.kurtosis	Y Probability Distribution
30	Number of Peaks	ela_distr.number_of_peaks	Y Probability Distribution
31	MMCE-LDA-10	ela_level.mmce_lda_10	Level Set
32	MMCE-QDA-10	ela_level.mmce_qda_10	Level Set
33	MMCE-MDA-10	ela_level.mmce_mda_10	Level Set
34	LDA-QDA-10	ela_level.lda_qda_10	Level Set
35	LDA-MDA-10	ela_level.lda_mda_10	Level Set
36	QDA-MDA-10	ela_level.qda_mda_10	Level Set
37	MMCE-LDA-25	ela_level.mmce_lda_25	Level Set
38	MMCE-LDA-10	ela_level.mmce_qda_25	Level Set
39	MMCE-MDA-25	ela_level.mmce_mda_25	Level Set
40	LDA-QDA-25	ela_level.lda_qda_25	Level Set
41	LDA-MDA-25	ela_level.lda_mda_25	Level Set
42	QDA-MDA-25	ela_level.qda_mda_25	Level Set
43	MMCE-LDA-50	ela_level.mmce_lda_50	Level Set
44	MMCE-QDA-50	ela_level.mmce_qda_50	Level Set
45	MMCE-MDA-50	ela_level.mmce_mda_50	Level Set
46	LDA-QDA-50	ela_level.lda_qda_50	Level Set
47	LDA-MDA-50	ela_level.lda_mda_50	Level Set
48	QDA-MDA-50	ela_level.qda_mda_50	Level Set
49	Linear Model w/o Interaction R2 Adj	ela_meta.lin_simple.adj_r2	Meta Model
50	Linear Model Intercept	ela_meta.lin_simple.intercept	Meta Model
51	Linear Model Minimum Coefficient	ela_meta.lin_simple.coef.min	Meta Model
52	Linear Model Maximum Coefficient	ela_meta.lin_simple.coef.max	Meta Model

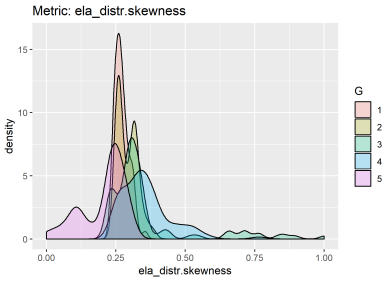
53	Linear Model Ratio Max-Min Coefficient	ela_meta.lin_simple.coef.max_by_min	Meta Model
54	Linear Model w/ Interaction R2 Adj	ela_meta.lin_w_interact.adj_r2	Meta Model
55	Quadratic Model w/o Interaction R2 Adj	ela_meta.quad_simple.adj_r2	Meta Model
56	Quadratic Model w/o Interaction Ratio Max-Min Quad. Coefficients	ela_meta.quad_simple.cond	Meta Model
57	Quadratic Model w/ Interaction R2 Adj	ela_meta.quad_w_interact.adj_r2	Meta Model
58	Absolute Number of Local Optima Clusters	ela_local.n_loc_opt.abs	Local Search
59	Relative Number of Local Optima Clusters	ela_local.n_loc_opt.rel	Local Search
60	Location of Global Optimum / Mean Location Local Optima	ela_local.best2mean_contr.orig	Local Search
61	Normalized Location of Global Optimum / Mean Location Local Optima	ela_local.best2mean_contr.ratio	Local Search
62	Average Size of Local Optima Cluster	ela_local.basin_sizes.avg_best	Local Search
63	Average Size of Non Local Optima	ela_local.basin_sizes.avg_non_best	Local Search
64	Average Size of Cluster of Worst Points	ela_local.basin_sizes.avg_worst	Local Search
65	Minimum Value of Local Optima	ela_local.fun_evals.min	Local Search
66	Lower Quartile of Values of Local Optima	ela_local.fun_evals.lq	Local Search
67	Mean Values of Local Optima	ela_local.fun_evals.mean	Local Search

68	Median Values of Local Optima	ela_local.fun_-evals.median	Local Search
69	Upper Quartile of Values of Local Optima	ela_local.fun_evals.uq	Local Search
70	Maximum Value Local Otima	ela_local.fun_-evals.max	Local Search
71	Std. Dev Value Local Optima	ela_local.fun_evals.sd	Local Search
72	Ratio Dispersion Mean Quantile 2%	disp.ratio_mean_02	Dispersion
73	Ratio Dispersion Mean Quantile 5%	disp.ratio_mean_05	Dispersion
74	Ratio Dispersion Mean Quantile 10%	disp.ratio_mean_10	Dispersion
75	Ratio Dispersion Mean Quantile 25%	disp.ratio_mean_25	Dispersion
76	Ratio Dispersion Median Quantile 2%	disp.ratio_median_02	Dispersion
77	Ratio Dispersion Median Quantile 5%	disp.ratio_median_05	Dispersion
78	Ratio Dispersion Median Quantile 10%	disp.ratio_median_10	Dispersion
79	Ratio Dispersion Median Quantile 25%	disp.ratio_median_25	Dispersion
80	Difference Dispersion Mean Quantile 2%	disp.diff_mean_02	Dispersion
81	Difference Dispersion Mean Quantile 5%	disp.diff_mean_05	Dispersion
82	Difference Dispersion Mean Quantile 10%	disp.diff_mean_10	Dispersion
83	Difference Dispersion Mean Quantile 25%	disp.diff_mean_25	Dispersion
84	Difference Dispersion Median Quantile 2%	disp.diff_median_02	Dispersion
85	Difference Dispersion Median Quantile 5%	disp.diff_median_05	Dispersion

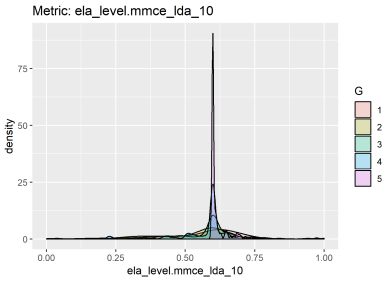
86	Difference Dispersion Median Quantile 10%	disp.diff_median_10	Dispersion
87	Difference Dispersion Median Quantile 25%	disp.diff_median_25	Dispersion
88	Maximum Entropy of Information Content	ic.h.max	Information Content
89	Epsilon Settling Sensitivity	ic.eps.s	Information Content
90	Maximum Epsilon	ic.eps.max	Information Content
91	Epsilon Ratio	ic.eps.ratio	Information Content
92	Partial Information Content	ic.m0	Information Content
93	Local Fitness Distance Correlation	FDC_local	-
94	Global Fitness Distance Correlation	FDC_Global	-
95	Correlation Length	Corr.Length	-
96	Y Number of Peaks	n_peaks	-
97	Y Skewness	skewness	-
98	Y Kurtosis	kurtosis	-
99	Y Median	median value	-
100	Lenght Scale Median	Length Scale Median	Length Scale
101	Length Scale Mode	mode	Length Scale
102	Length Scale Variance	var	Length Scale
103	Length Scale Maximum	max	Length Scale
104	Length Scale Minimum	min	Length Scale
105	Length Scale Kernel Bandwidth	bandwidth	Length Scale
106	Length Scale Entropy	entropy	Length Scale

Appendix E - Density Plots

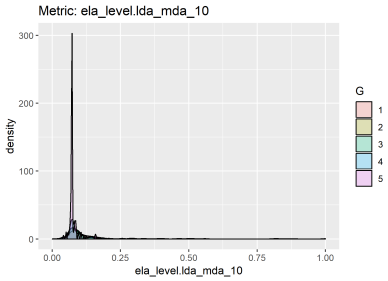
E.1 Density Plots



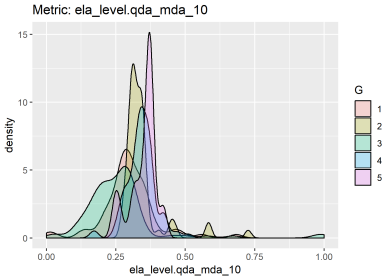
(a) Skewness



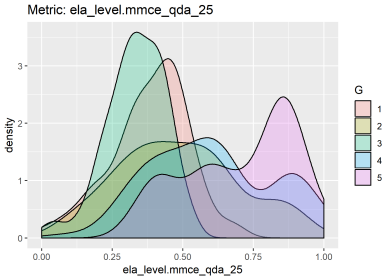
(b) MMCE-LDA-10



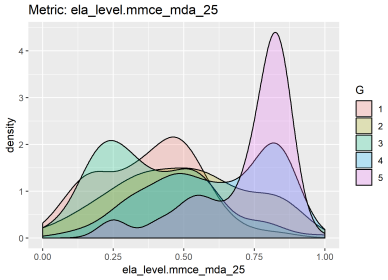
(c) LDA-MDA-10



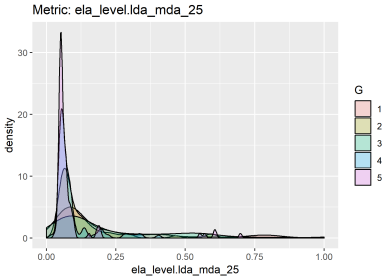
(d) QDA-MDA-10



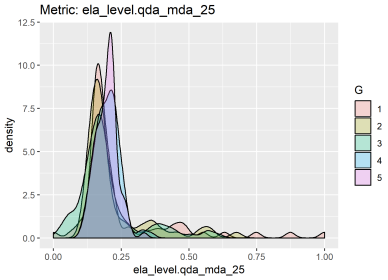
(e) MMCE-QDA-25



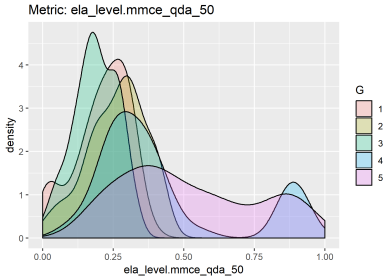
(f) MMCE-MDA-25



(g) LDA-MDA-25

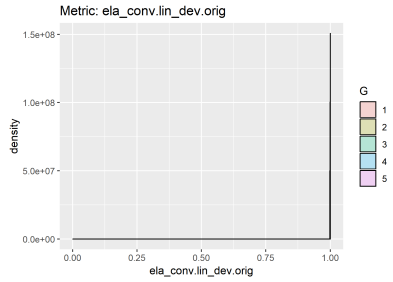


(h) QDA-MDA-25

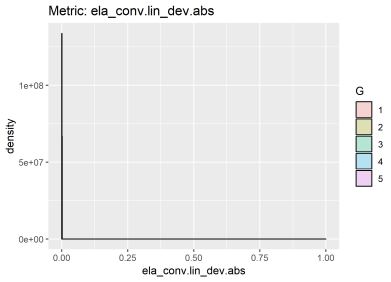


(i) MMCE-QDA-50

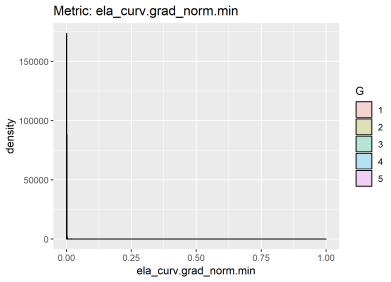
FIGURE E.1 – ELA Metrics in Set 1 - A



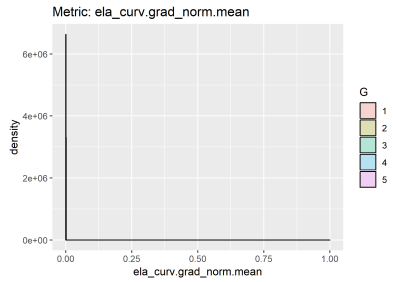
(a) Linear Deviation to Origin



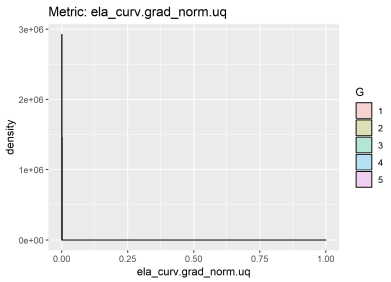
(b) Absolute Linear Deviation



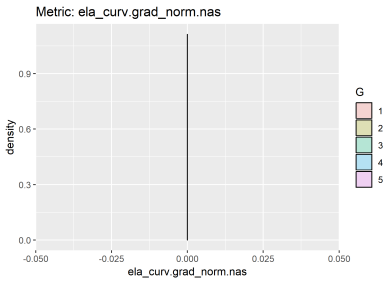
(c) Minimum of Norm. Gradient



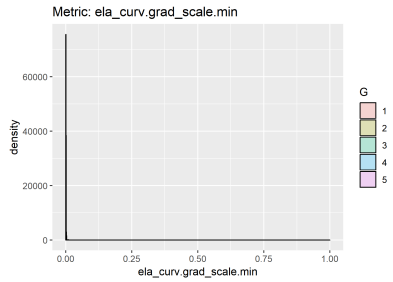
(d) Mean of Norm. Gradient



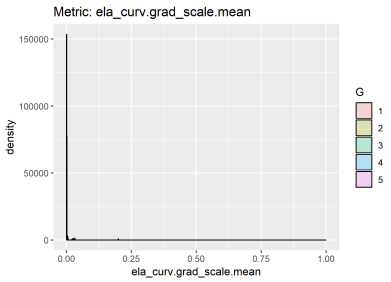
(e) Upper Qrt. of Norm. Gradient



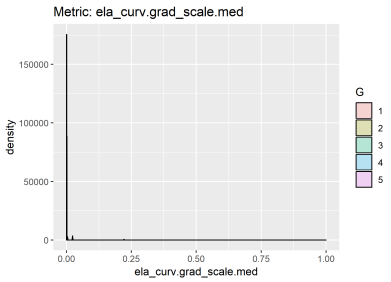
(f) N. of NAs in Norm. Gradient



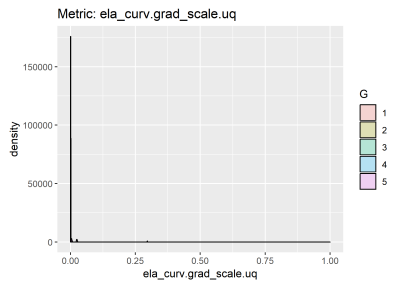
(g) Minimum of Scaled Gradient



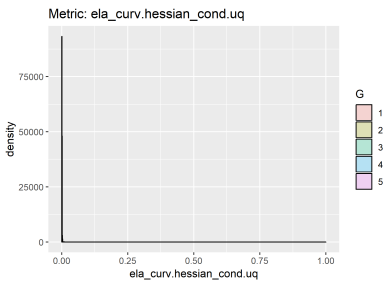
(h) Mean of Scaled Gradient



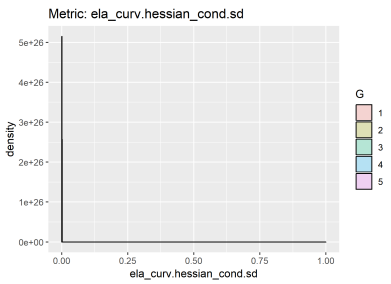
(i) Median of Scaled Gradient



(j) Upper Qrt. Scaled Gradient

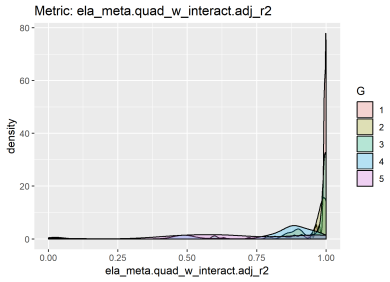


(k) Upp. Qrt. Hessian Matrix Cond.

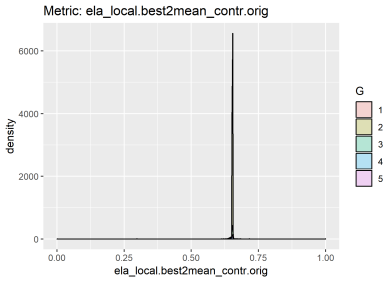


(l) Upp. Qrt. Hessian Matrix Std.Dev

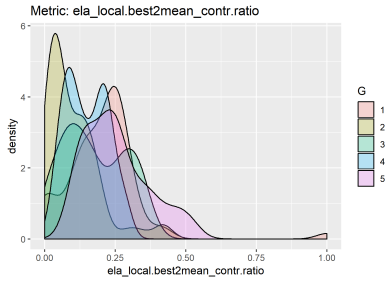
FIGURE E.2 – ELA Metrics in Set 1 - B



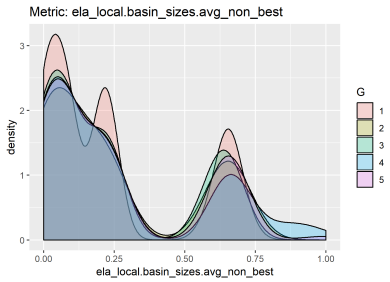
(a) Quad. with Interaction Adj-R2



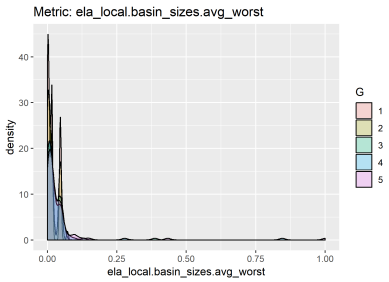
(b) Best to Mean Contrast Org.



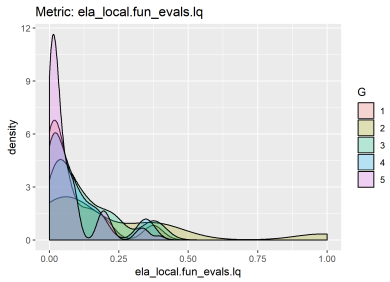
(c) Best to Mean Contrast Ratio



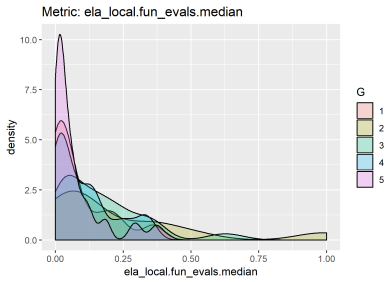
(d) Basin Sizes Avg. Non Best



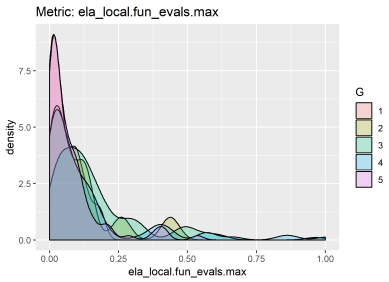
(e) Basin Sizes Avg. Worst



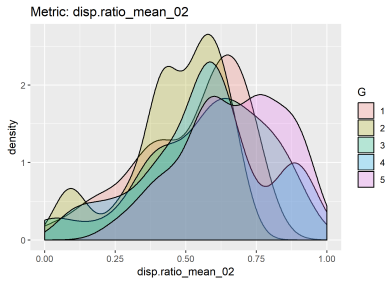
(f) Lower Qrt. Function Evaluation



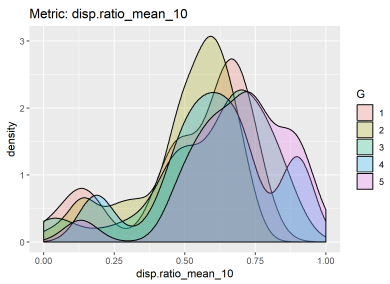
(g) Median Function Evaluation



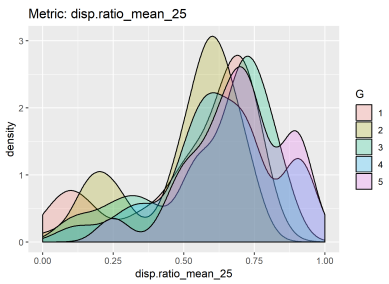
(h) Maximum Function Evaluation



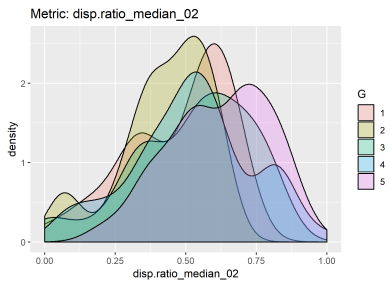
(i) Mean Dispersion Ratio 2%



(j) Mean Dispersion Ratio 10%

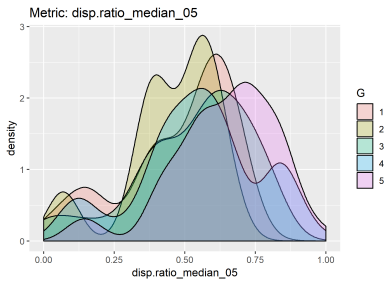


(k) Mean Dispersion Ratio 25%

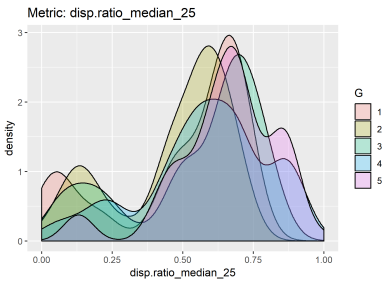


(l) Median Dispersion Ratio 2%

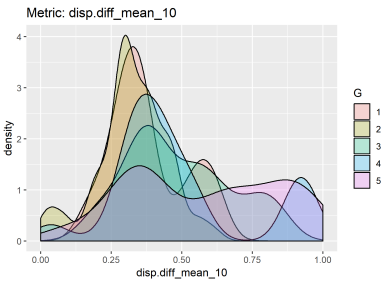
FIGURE E.3 – ELA Metrics in Set 1 - C



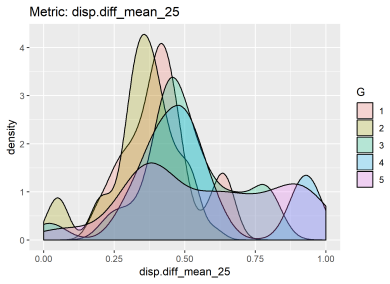
(a) Median Dispersion Ratio 5%



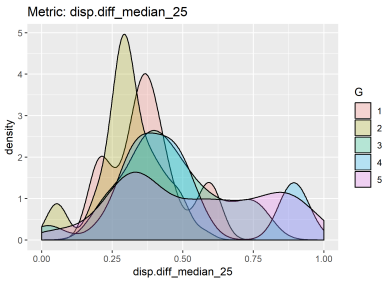
(b) Median Dispersion Ratio 25%



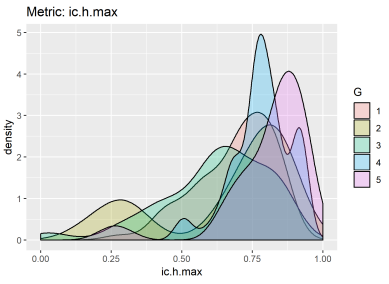
(c) Mean Dispersion Diff. 10%



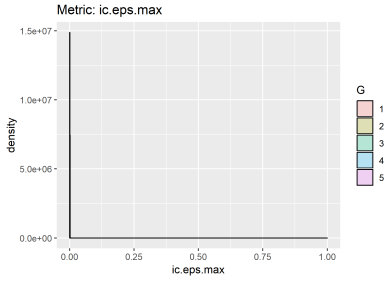
(d) Mean Dispersion Diff. 25%



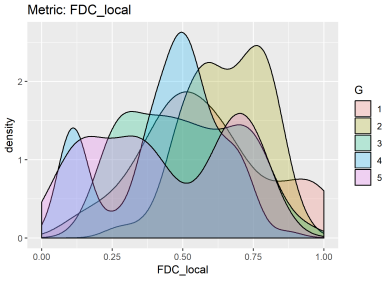
(e) Median Dispersion Diff. 25%



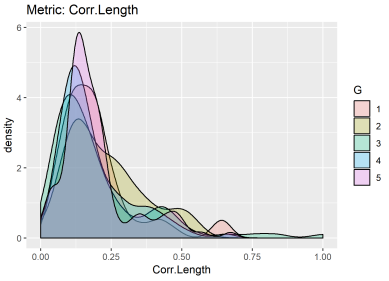
(f) Max Entropy



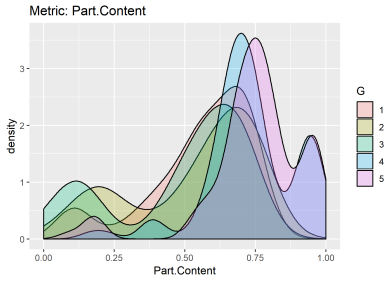
(g) Max Epsilon



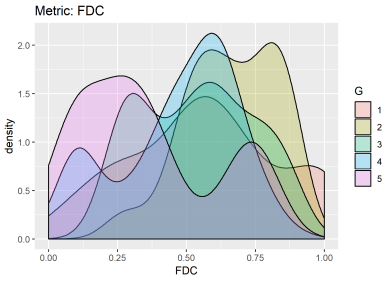
(h) Local FDC



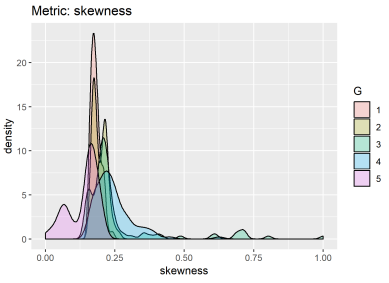
(i) Correlation Length



(j) Partial Info. Content

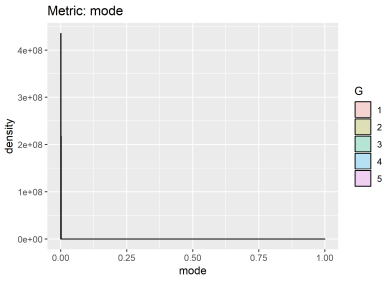


(k) Global FDC

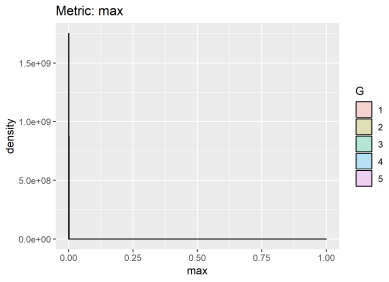


(l) Length Scale Skewness

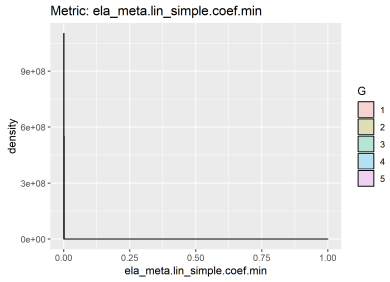
FIGURE E.4 – ELA Metrics in Set 1 - D



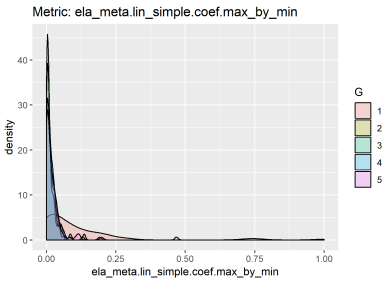
(a) Mode Length Scale

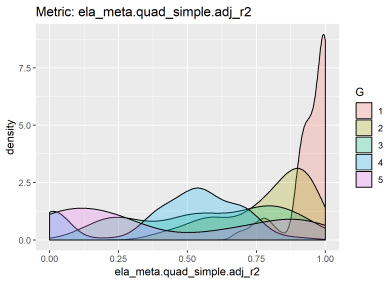


(b) Maximum Length Scale



(c) Coefficients Linear Model



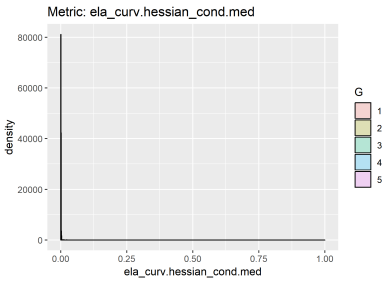


(d) Linear Simple Coeffs. Max-Min (e) Quadratic Adjusted R2 Simple

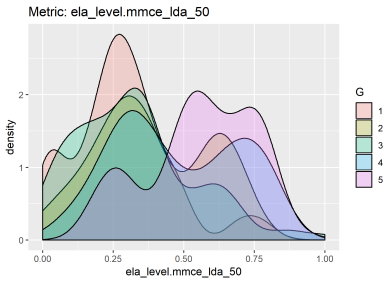
FIGURE E.5 – ELA Metrics in Set 1 - E

E.1.0.1 Proposed Set 2 - ELA 10 Metrics

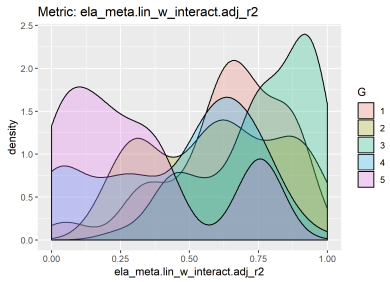
Figures E.6 and E.7 depict the density plots of metrics included in the proposed Set 2, which contains 10 ELA metrics.



(a) Median Hessian Matrix Cond.

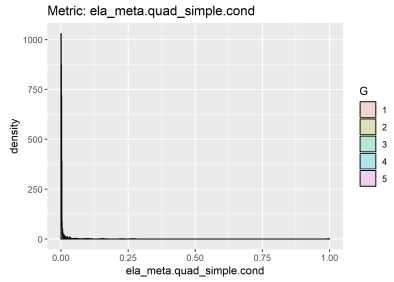


(b) MMCE-LDA-50

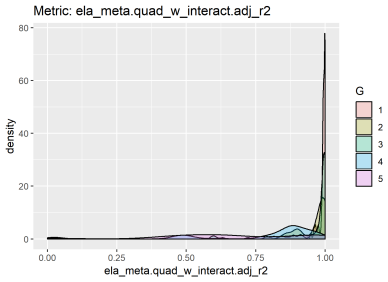


(c) Linear with Interaction Adj-R2

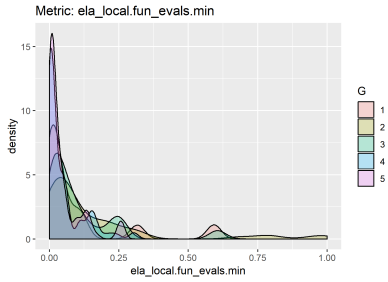
FIGURE E.6 – ELA Metrics in Set 2 - A



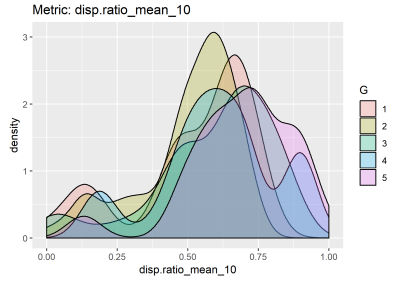
(a) Quadratic Model Condition



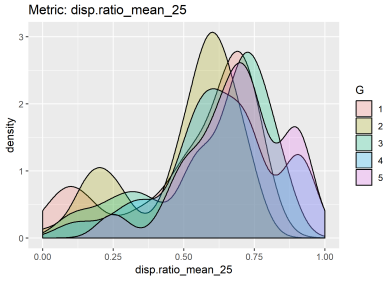
(b) Quadratic with Interaction Adj-R2



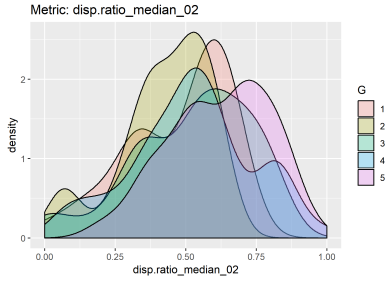
(c) Min Function Evaluations



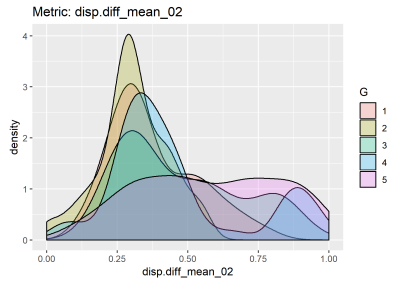
(d) Mean Dipersion Ratio 10%



(e) Mean Dispersion Ratio 25%



(f) Median Dispersion Ratio 2%

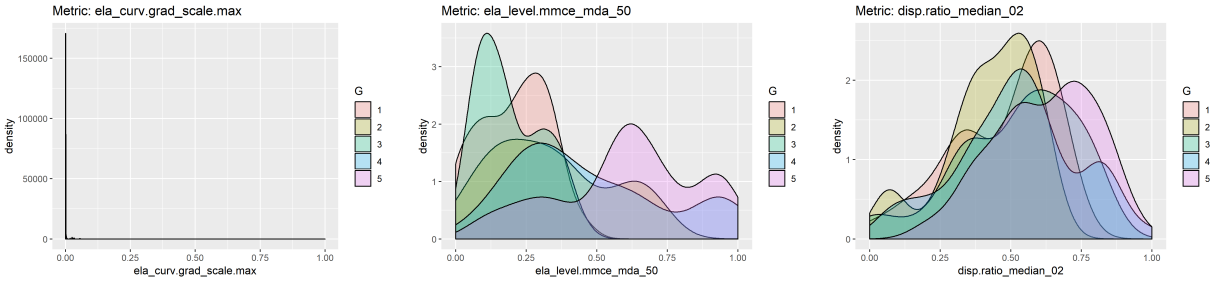


(g) Mean Dispersion Diff. 2%

FIGURE E.7 – ELA Metrics in Set 2 - B

E.1.0.2 Proposed Set 3 - 3 ELA Metrics

Figure E.8 depicts the density plot of ELA metrics included in set 3.

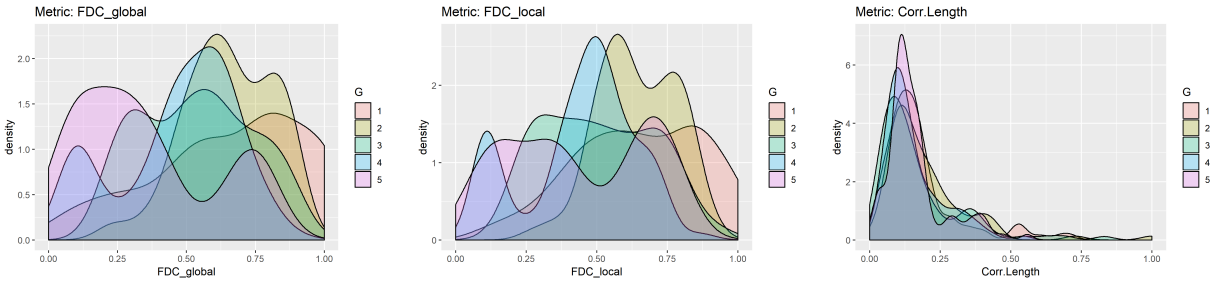


(a) Maximum Gradient Scaled (b) MMCE-MDA 50 (c) Median Dispersion Ratio 2%

FIGURE E.8 – ELA Metrics in Set 3 - A

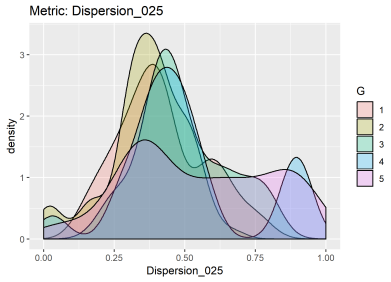
E.1.0.3 Set M - 7 ELA Metrics

Figures E.9 and E.10 depict the density plot of ELA metrics included in set M as in (MORGAN; GALLAGHER, 2017).

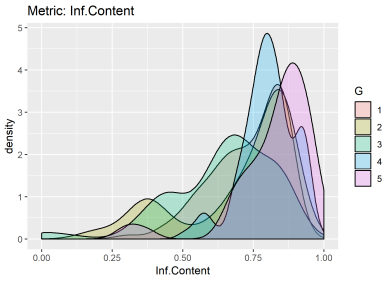


(a) FDC Global (b) FDC Local (c) Correlation Length

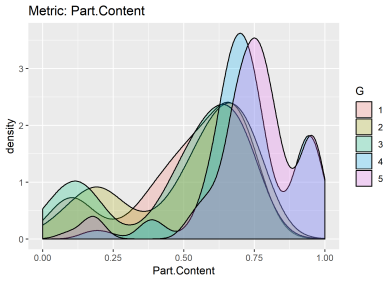
FIGURE E.9 – ELA Metrics in Set M - A



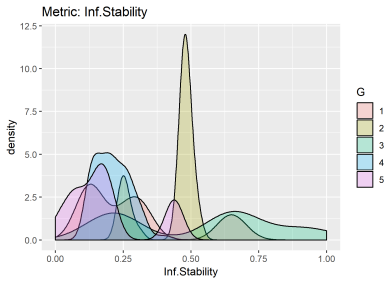
(a) Median Dispersion Ratio 25%



(b) Info. Content



(c) Partial Info. Content

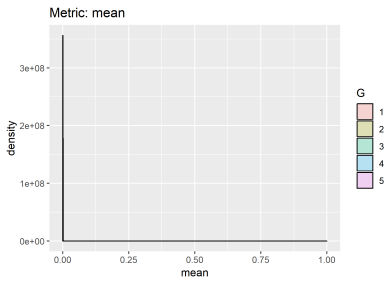


(d) Info. Stability

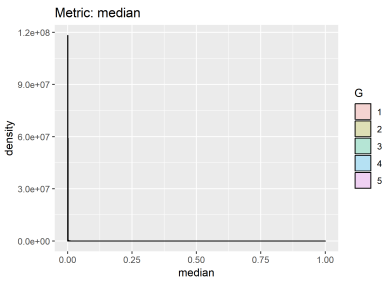
FIGURE E.10 – ELA Metrics in Set M - B

E.1.0.4 Set LS - 8 ELA Metrics

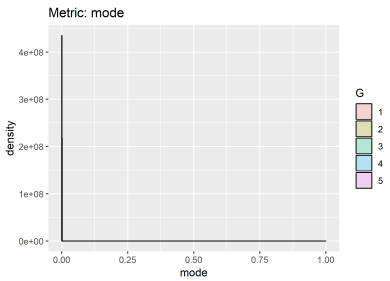
Figures E.11 and E.12 depict the density plot of ELA metrics included in set LS as in (MORGAN; GALLAGHER, 2017).



(a) Mean Length Scale

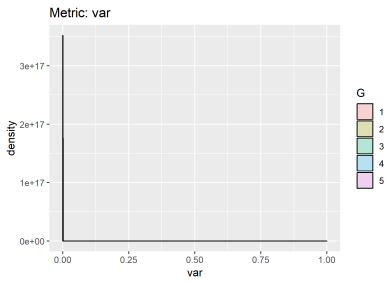


(b) Median Length Scale

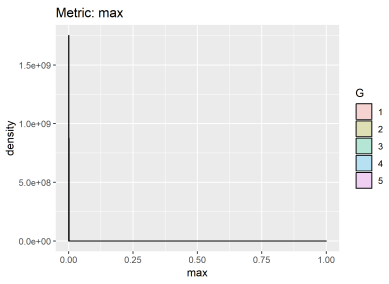


(c) Mode Length Scale

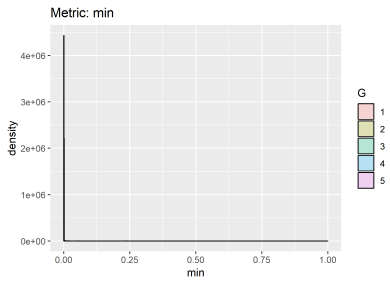
FIGURE E.11 – ELA Metrics in Set LS - A



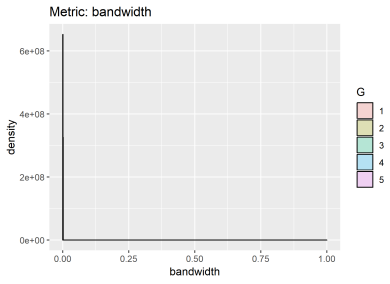
(a) Variance Length Scale



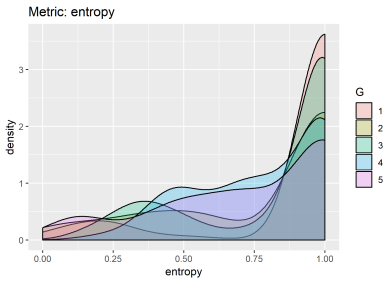
(b) Maximum Length Scale



(c) Minimum Length Scale



(d) LS Kernel Bandwidth

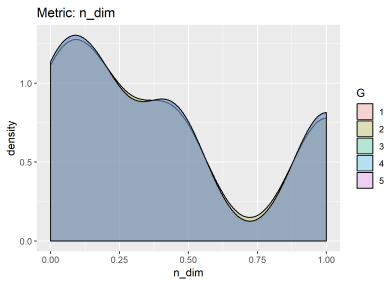


(e) Length Scale Entropy

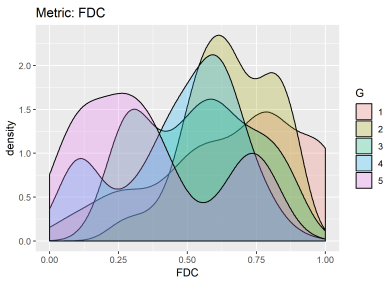
FIGURE E.12 – ELA Metrics in Set LS - B

E.1.0.5 Set SU - 9 Metrics

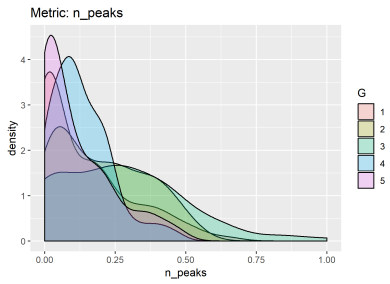
Figures E.13 and E.14 depict the density plot of ELA metrics included in set SU as in (SUN *et al.*, 2014).



(a) Number of Dimensions

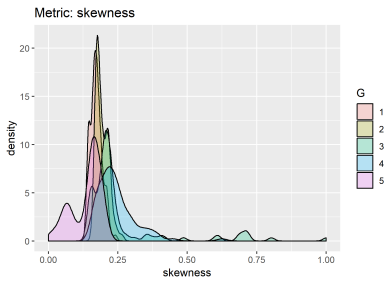


(b) FDC Global

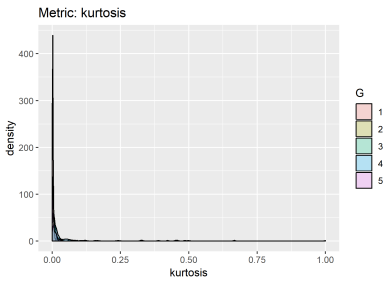


(c) Number of Peaks

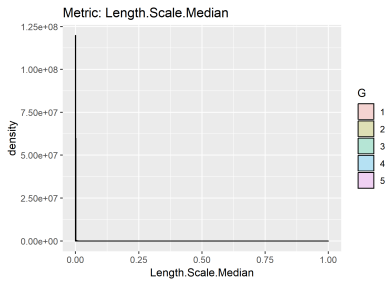
FIGURE E.13 – ELA Metrics in Set SU - A



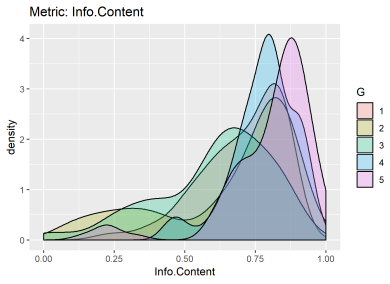
(a) Skewness



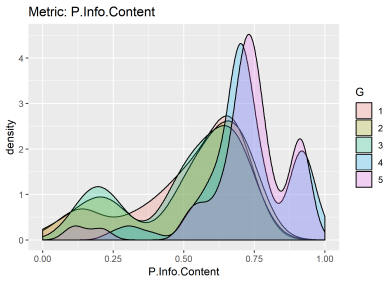
(b) Kurtosis



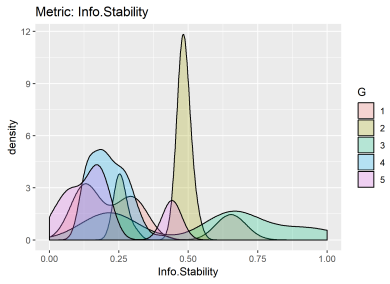
(c) Median Length Scale



(d) Information Content



(e) Partial Info. Content

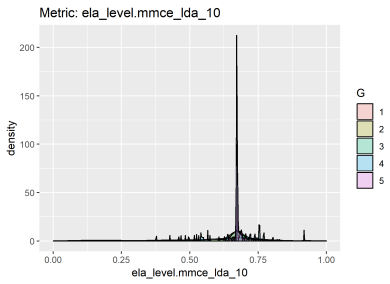


(f) Information Stability

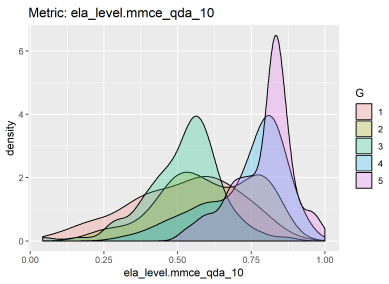
FIGURE E.14 – ELA Metrics in Set SU - B

E.1.0.6 Set MER - 18 Metrics

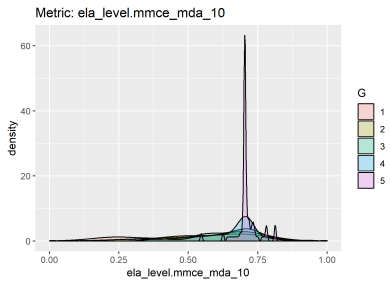
Figures E.15, E.16 and E.17 depict the density plot of ELA metrics included in set MER as in (MERSMANN *et al.*, 2010).



(a) MMCE-LDA-10

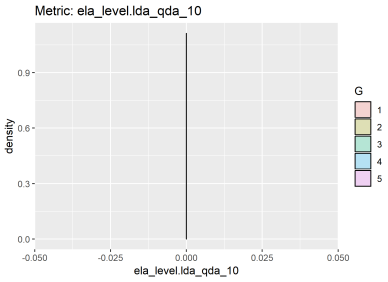


(b) MMCE-QDA-10

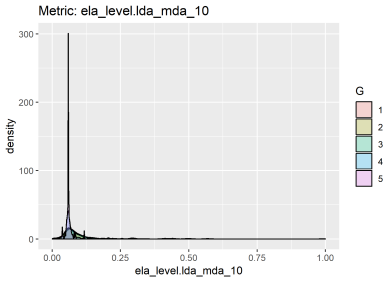


(c) MMCE-MDA-10

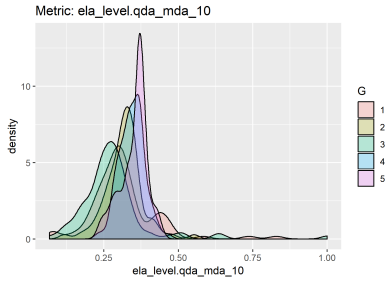
FIGURE E.15 – ELA Metrics in Set MER - A



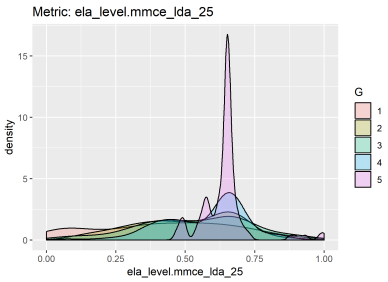
(a) LDA-QDA-10



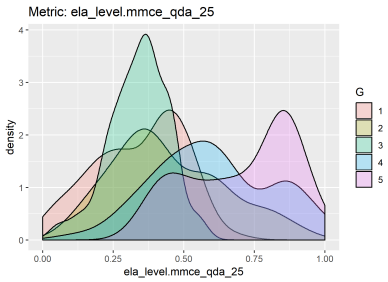
(b) LDA-MDA-10



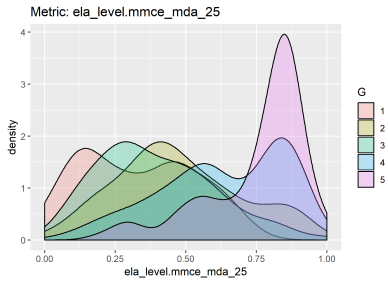
(c) QDA-MDA-10



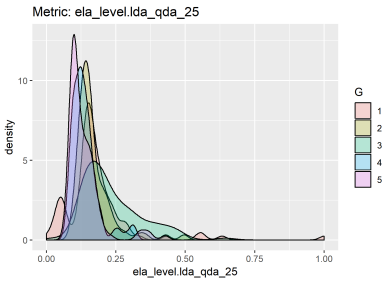
(d) MMCE-LDA-25



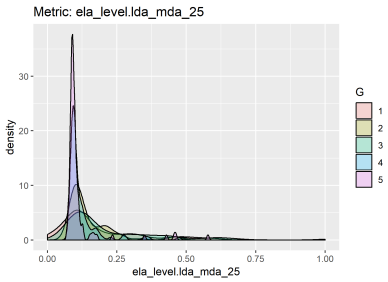
(e) MMCE-QDA-25



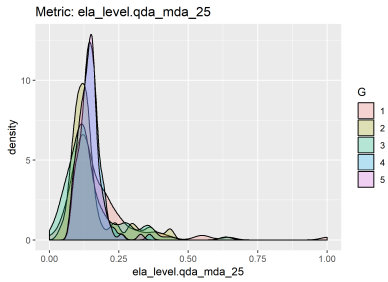
(f) MMCE-MDA-25



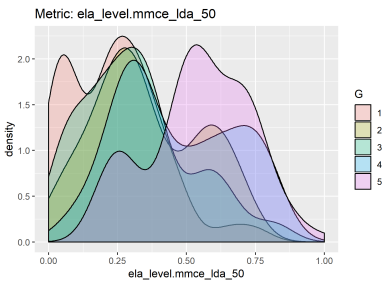
(g) LDA-QDA-25



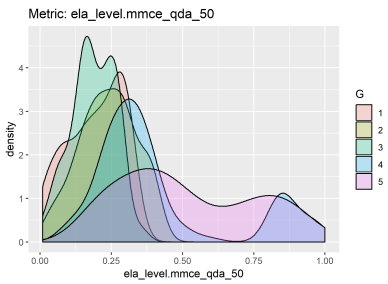
(h) LDA-MDA-25



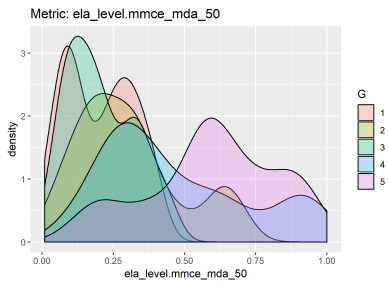
(i) QDA-MDA-25



(j) MMCE-LDA-50

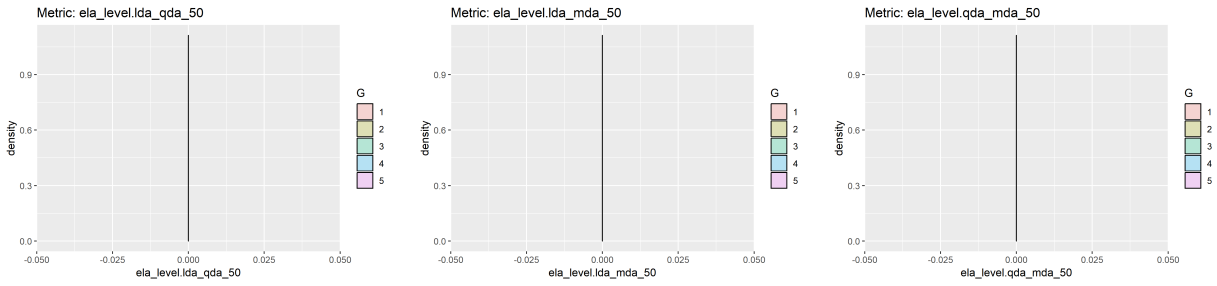


(k) MMCE-QDA-50



(l) MMCE-MDA-50

FIGURE E.16 – ELA Metrics in Set MER - B



(a) LDA-QDA-50 (b) LDA-MDA-50 (c) QDA-MDA-50

FIGURE E.17 – ELA Metrics in Set MER- C